THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

# $G$-Equivariant Convolutional Neural Networks

JIMMY ARONSSON


CHALMERS

Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden 2021

$G$-Equivariant Convolutional Neural Networks
*Jimmy Aronsson*



Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden
Telephone +46 (0)31 772 1000

Author e-mail: jimmyar@chalmers.se

Typeset with LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

# $G$-Equivariant Convolutional Neural Networks

**Jimmy Aronsson**

Division of Algebra and Geometry
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

## Abstract

Over the past decade, deep learning has revolutionized industry and academic research. Neural networks have been used to solve a multitude of previously unsolved problems and to significantly improve the state-of-the-art on other tasks, in some cases reaching superhuman levels of performance. However, most neural networks have to be carefully adapted to each application and often require large amounts of data and computational resources.

*Geometric deep learning* aims to reduce the amount of information that neural networks have to learn, by taking advantage of geometric properties in data. In particular, *equivariant neural networks* use (local or global) symmetry to reduce the complexity of a learning task.

In this thesis, we investigate a popular deep learning model for tasks exhibiting global symmetry: *G-equivariant convolutional neural networks (GCNNs)*. We analyze the mathematical foundations of GCNNs and discuss where this model fits in the broader scheme of equivariant learning. More specifically, we discuss a general framework for equivariant neural networks using notions from gauge theory, and then show how GCNNs arise from this framework in the presence of global symmetry. We also characterize *convolutional layers*, the main building blocks of GCNNs, in terms of more general *G-equivariant layers* that preserve the underlying global symmetry.

**Keywords:** deep learning, convolutional neural networks, homogeneous spaces, homogeneous vector bundles, induced representations, symmetry.

## List of publications

This thesis is based on the work represented by the following paper:

I. **Aronsson, J.** (2021). Homogeneous vector bundles and $G$-equivariant convolutional neural networks. *arXiv preprint arXiv:2105.05400.*

Additional papers not included in this thesis:

II. Gerken, J.E., **Aronsson, J.**, Carlsson, O., Linander, H., Ohlsson, F., Petersson, C., Persson, D. (2021). Geometric Deep Learning and Equivariant Neural Networks. *To appear*.

## Author contributions

II. Foundational work on gauge equivariant neural networks and GCNNs, including detailed mathematical analysis and ensuring mathematical rigor. Main responsible for the section on GCNNs.

# Acknowledgements

First and foremost I would like to thank my supervisor Daniel Persson, my co-supervisors Christoffer Petersson and Robert Berman, and all other members of my research group: Oscar Carlsson, Jan E. Gerken, Hampus Linander and Fredrik Ohlsson. This venture into geometric deep learning has truly gained momentum and it is a joy to work together with you.

Thank you to the members of AI-Batch 1 of the Wallenberg AI and Autonomous Systems and Software Program (WASP), and to my friends at the Department of Mathematical Sciences: Carl, Carl-Joar, Edvin, Felix, Gabrijela, Gustav, Jimmy, Juan, Kristian, Linnea, Linnea, Malin, Olof, Oskar, Per, and anyone else I may have forgotten to mention. Thank you also to Aila Särkkä, Elisabeth Eriksson, Håkan Samuelsson-Kalm, Marie Kühn, and others for helping me and patiently answering my many questions about life at the Department of Mathematical Sciences.

Simultaneously working on this thesis and on two research papers has been incredibly stressful. I could not have done it without the love and support from my girlfriend Iris. I am in awe of you ability to focus only on things that give meaning to life, and I have never before laughed as much as I do with you. Finally, a special mention to Mrs Agapi Aronsson-Dog, the most anxious but loving dog I have ever met.

# Contents

**Abstract** **iii**

**List of publications** **v**

**Acknowledgements** **vii**

**Contents** **ix**

**1 Introduction** **1**

1.1 Deep Learning . . . . . 2

1.2 Representation theory . . . . . 10

1.3 Fiber bundles . . . . . 18

**2 Summary of results** **25**

**Bibliography** **27**

**Paper I**

# 1 Introduction

In this chapter, we provide the relevant background material that is needed in order to understand **Paper I**.

First, we give a basic introduction to deep learning. This introduction focuses on a certain class of neural networks and a specific learning task, but its contents are much more widely applicable. We then delve into more detailed discussions on convolutional neural networks and geometric deep learning.

Geometric deep learning models, in particular the *G-equivariant convolutional neural networks* and *gauge equivariant neural networks* studied in **Paper I**, use the language of principal bundles and associated vector bundles to describe learning on manifolds. In turn, these notions require basic knowledge of group representations. We therefore give short introductions to representation theory and the theory of fiber bundles.

## 1.1 Deep Learning

We begin with an introduction to deep learning, and then discuss convolutional neural networks and geometric deep learning. See Goodfellow et al. (2016) for a more comprehensive introduction to deep learning.

### 1.1.1 Essentials of deep learning

Consider the problem of learning an unknown function $f : \mathcal{X} \to \mathcal{Y}$ between two spaces $\mathcal{X}$ and $\mathcal{Y}$, given a *training data set*

$$S_{\text{train}} = \{(x_i, f(x_i)) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \dots, N_{\text{train}}\}. \tag{1.1}$$

In a house price estimation task, for instance, $\mathcal{X}$ could be the space of all houses in a given geographical region, parameterized by numerical features such as postal codes, area in square meters, number of rooms, age, etc. In this example, the codomain $\mathcal{Y} = [0, M]$ would be the possible price range, for some realistic upper bound $M > 0$ on the price, and the unknown function $f : \mathcal{X} \to \mathcal{Y}$ would associate each house with its, in some sense, correct price. The training data set $S_{\text{train}}$ would contain a list of recently sold houses $x_i$ in the given geographical region together with the final selling price $f(x_i)$.

Another example is an image classification task in which $\mathcal{X}$ contains satellite images and $\mathcal{Y} = \{1, 2, 3, \dots, k\}$ is a list of countries. Perhaps $y = 1$ represents Sweden, $y = 2$ represents Denmark, etc. The unknown function $f : \mathcal{X} \to \mathcal{Y}$ would then map each image to the country that it depicts. There could also be separate classes $y \in \mathcal{Y}$ for edge cases such as country borders, oceans, and so on. In this example, the training set $S_{\text{train}}$ would be a relatively small subset of images $x_i$ for which the depicted country $f(x_i)$ is known.

A natural approach to approximating such unknown functions is to consider a space of parameterized functions $f_\theta : \mathcal{X} \to \mathcal{Y}$, and use the training data set to optimize the parameter $\theta$ so to minimize the discrepancy between $f_\theta$ and $f$. This is what neural networks do.

Among the simplest and most basic neural networks are *multilayer perceptrons*. These networks are functions $f_\theta : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ consisting of a sequence of *layers*

$$x^l = \sigma^l\left(W^l x^{l-1} + b^l\right), \qquad l = 1, \dots, L, \tag{1.2}$$

where $x^0 \in \mathbb{R}^{N_0}$ is the input to the first layer. In each layer $l = 1, \dots, L$, the components of the *bias vector* $b^l \in \mathbb{R}^{N_l}$ and of the *weight matrix* $W^l \in \mathbb{R}^{N_l \times N_{l-1}}$

are so-called *trainable parameters*, and the *activation function* $\sigma^l : \mathbb{R} \to \mathbb{R}$ is a non-linear function that is applied element-wise to $W^l x^{l-1} + b^l \in \mathbb{R}^{N_\ell}$. There are a few common choices for the activation function, mainly the rectified linear unit $\mathrm{ReLU}(x) = \max\{0, x\}$ and the sigmoid $\sigma(x) = (1 + e^{-x})^{-1}$. To summarize, the neural network is the function

$$f_\theta : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, \qquad f_\theta(x^0) = x^L, \tag{1.3}$$

and $\theta$ represents all components of the bias vectors $b^l$ and weight matrices $W^l$. The number of parameters can thus grow extremely large. The dimensions $N_0$ and $N_L$ are chosen under the assumption that $\mathcal{X} = \mathbb{R}^{N_0}$ and $\mathcal{Y} = \mathbb{R}^{N_L}$.

As stated above, we can use the training data to compare $f_\theta$ with the unknown function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$, by comparing the output $f_\theta(x_i)$ of the neural network with its corresponding *label* $y_i = f(x_i)$, for each training data point $x_i$. To make the comparison, we use a non-negative *loss function*

$$\ell : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to [0, \infty), \tag{1.4}$$

common choices being the Euclidean norm $\ell(v, w) = \|v - w\|$ or its square $\ell(v, w) = \|v - w\|^2$. We compute the loss for all elements $(x_i, y_i) \in S_{\text{train}}$ of the training data set and consider the minimization problem

$$\min_\theta \underbrace{\sum_i \ell(f_\theta(x_i), y_i)}_{\ell(\theta)}. \tag{1.5}$$

The neural network $f_\theta$ is *trained* by iteratively improving the parameter $\theta$, i.e., the biases and the weight matrices, until the *training loss* $\ell(\theta)$ is sufficiently small. The trainable parameters in $\theta$ are often initialized randomly, hence the neural network produces nonsense estimates in the beginning. The network has no predictive power in the initial stages. However, an iterative optimization method such as gradient descent,

$$\theta \mapsto \theta - \alpha \nabla \ell(\theta), \tag{1.6}$$

with *learning rate* $0 < \alpha < 1$, ensures that the training loss $\ell(\theta)$ always decreases in each iteration of training (called an *epoch* to distinguish it from other iterative processes). There are many different optimization methods to choose between, but most methods are based on gradient descent. Alternatively, using *stochastic* gradient descent allows the training loss to occasionally increase, thus reducing the risk of getting stuck in a local minima, but the loss still decreases over time.

Given sufficiently large numbers of epochs and trainable parameters, a well-designed neural network eventually reaches a small training loss. The network can then accurately predict the correct labels $y_i$ for the training data points $x_i$. This does not imply, however, that the performance generalizes beyond the training data. Perhaps the network simply learned the data by heart, instead of learning relevant features that would help it make good predictions $f_\theta(x)$ for data points $x$ that it has not previously seen. If so, the neural network is said to have *overfit* the training data.

One example would be if there is only a single training data point, $N_{\text{train}} = 1$. The network has little chance to learn general features of the data distribution if it only sees a single instance, hence it is likely to overfit. One way to prevent overfitting is thus to increase the amount of training data. Another method is to add regularization terms to (1.5) that prevent the training loss from becoming too small.

The performance is also evaluated during each epoch, each iteration of training, by applying the network to a separate *test data* set $S_{\text{test}}$ and computing the *test loss*. This information is not used directly to update the network parameters via gradient descent, it is only for evaluation. That being said, performance on test data is often used as a stopping criteria: If the test loss begins to increase while the training loss is still decreasing, the network is likely starting to overfit. Training may thus be aborted at this point.

Model parameter such as the number of layers $L$, the dimension $N_l$ of each layer, regularization parameters, and so on, are referred to as *hyperparameters*. It is common to train a neural network multiple times with different combinations of hyperparameters, and then choose the best performing combination. The test data has then indirectly influenced the training of all these network designs, which introduces bias, so a third *validation* data set $S_{\text{validation}}$ is used to evaluate the different combinations of hyperparameters. Hopefully, the resulting neural network $f_\theta$ is a good approximation to the unknown function $f$.

In this introduction to deep learning, we have discussed one of the most basic types of neural networks - the multilayer perceptron. We have also focused on the task of approximating an unknown function $f : \mathcal{X} \to \mathcal{Y}$ given a training data set (1.1) containing labeled data $(x_i, y_i)$. That is, a *supervised* learning task. While there are a multitude of different neural networks and different kinds of tasks, with different training procedures and evaluation methods, the material we have presented summarizes some essential aspects of deep learning.

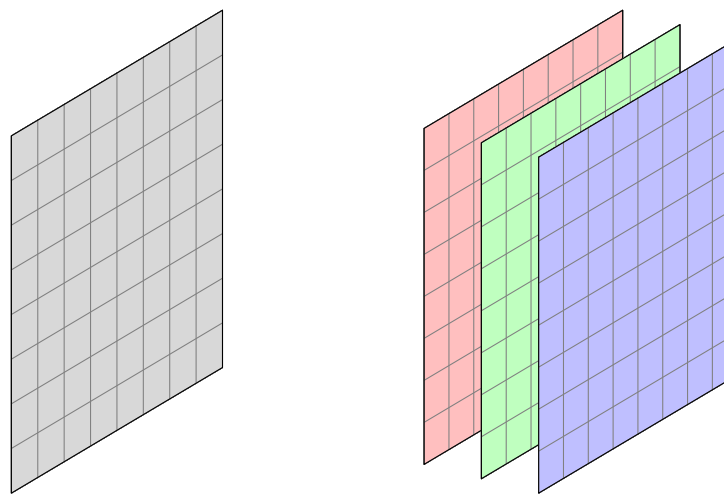
### 1.1.2 Convolutional neural networks

In this thesis, we are primarily interested in a particular type of neural networks called *convolutional neural networks (CNNs)*. These networks are applied to *data points* that have a 2D or 3D grid structure, which we represent mathematically as finitely supported functions

$$f : \mathbb{Z}^2 \to \mathbb{R}^m, \tag{1.7}$$

where $m$ is the number of *channels*. In this case, 3D grids correspond to $m > 1$ channels when thinking of the channels as the third grid dimension.

Digital images satisfy (1.7) if we view $\mathbb{Z}^2$ as the pixel grid: Digital images map each pixel to either a grayscale value ($m = 1$) or an RGB array ($m = 3$). Finite support is then analogous to finite image resolution. Equivalently, we can view RGB images as having a 3D grid structure in which the red, green, and blue color channels are stacked on top of each other (Figure 1.1).



**(a)** Grayscale image ($m = 1$). **(b)** RGB image ($m = 3$).

**Figure 1.1:** Digital images have a 2D (grayscale) or 3D (RGB) grid structure. We can also view them as functions $f : \mathbb{Z}^2 \to \mathbb{R}^m$ with $m = 1$ (grayscale) or $m = 3$ (RGB) channels.

The main building blocks of CNNs are called *convolutional layers*. The name is inspired by convolutions of real-valued functions $\kappa, f : \mathbb{R} \to \mathbb{R}$,

$$(\kappa * f)(y) = \int_{-\infty}^{\infty} \kappa(y - x) f(x) \, \mathrm{d}x, \tag{1.8}$$

but have been generalized to allow vector-valued data points (1.7):

$$[\kappa \star f](y) = \sum_{x \in \mathbb{Z}^2} \kappa(x - y) f(x). \tag{1.9}$$

Here, $\kappa : \mathbb{Z}^2 \to \mathrm{Hom}(\mathbb{R}^m, \mathbb{R}^n)$ is a matrix-valued *kernel*, or *filter*, for some natural number $n \in \mathbb{N}$. Note that (1.9) differs from (1.8) not only in dimensionality and the domain of integration/summation, we have also involuted the kernel: $\kappa(x - y)$ versus $\kappa(y - x)$. This means that (1.9) is a cross-correlation operator rather than a convolution operator, but it can easily be turned into the latter by redefining the kernel. **Paper I** uses the convention (1.9) since it makes some proofs easier to formulate.

Convolutional layers are trained by optimizing the matrix elements in $\kappa(x)$ for each $x \in \mathbb{Z}^2$. This is possible because, in practice, $\kappa$ is only supported on a small number of points around the origin in $\mathbb{Z}^2$. When referring to a convolutional layer with a "$3 \times 3$ kernel", for instance, we mean that

$$\mathrm{supp}(\kappa) = \big\{(x_1, x_2) \in \mathbb{Z}^2 \;\big|\; x_1, x_2 = -1, 0, 1\big\}. \tag{1.10}$$

That is, the kernel size $3 \times 3$ only refers to the support of $\kappa$, which is different from the matrix dimensions $n \times m$ of $\kappa(x) \in \mathrm{Hom}(\mathbb{R}^m, \mathbb{R}^n)$. The most common kernel sizes are $k \times k$ where $k$ is a small, odd integer.

To compute the output $[\kappa \star f](y)$ of a convolutional layer at a point $y \in \mathbb{Z}^2$, we first transform each $m$-channel input array $f(x)$ into an $n$-channel output array $\kappa(x - y) f(x)$ for each point $x \in \mathbb{Z}^2$, and then sum over $x$. This procedure can be visualized as placing a $k \times k$ kernel on top of the input data point $f$, with the kernel support centered at $y$, and computing pointwise inner products between $f$ and each row in the kernel $\kappa$ (Figure 1.2). When computing $[\kappa \star f](z)$ at another point $z \in \mathbb{Z}^2$, we simply reposition the kernel and repeat the process. Convolutional layers are thereby computed by "sliding" the kernel across the input data point $f$.

The "sliding kernel" interpretation illustrates that convolutional layers employ *weight sharing*, i.e., the same $k^2 * n * m$ non-zero kernel matrix elements are used to compute the output at each point. The very small number of weights in convolutional layers makes CNNs relatively efficient to train.

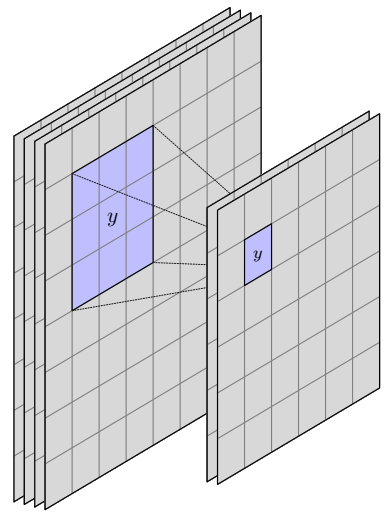



**Figure 1.2:** A convolutional layer that maps a 4-channel data point $f : \mathbb{Z}^2 \to \mathbb{R}^4$ into a 2-channel data point $\kappa \star f : \mathbb{Z}^2 \to \mathbb{R}^2$. This convolutional layer uses a $3 \times 3$ kernel, meaning that $\kappa$ is supported on a $3 \times 3$ grid. This should not be confused with the $4 \times 2$ matrix dimension of $\kappa(x)$.

There is another interesting consequence of the sliding kernel: Consider the translation operator on $\mathbb{Z}^2$ that translates each grid point by the same amount,

$$L_{x_0} : \mathbb{Z}^2 \to \mathbb{Z}^2, \qquad L_{x_0}(x) = x + x_0, \qquad x_0 \in \mathbb{Z}^2. \tag{1.11}$$

This translation operator induces a translation operator on data points $f$, that moves the argument in the opposite direction: $(L_{x_0} f)(x) = f(x - x_0)$. If we now apply a convolutional layer to the translated data point, we find that

$$\begin{aligned} [\kappa \star L_{x_0} f](y) &= \sum_{x \in \mathbb{Z}^2} \kappa(x - y) f(x - x_0) \\ &= \sum_{x \in \mathbb{Z}^2} \kappa(x - (y - x_0)) f(x) = L_{x_0}[\kappa \star f](y). \end{aligned} \tag{1.12}$$

Convolutional layers thus commute with the translation operator. Intuitively, this means that convolutional layers preserve the global symmetry in $\mathbb{Z}^2$, which is important for applications. As an example, consider using a CNN for a facial recognition task, and suppose for argument's sake that the kernel $\kappa$ has learned to recognize human eyes. Thanks to the *translation equivariance* (1.12), it does not matter where the eyes are located in any particular image, the sliding kernel

will eventually locate them. That is, the CNN does not have to worry too much about technical artifact such as the exact pixel coordinates of facial features.

We can add a bias vector $b \in \mathbb{R}^n$ after the convolutional layer and then apply a non-linear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ to each component. These operations are independent of $x \in \mathbb{Z}^2$, hence the composition

$$\sigma\left(\kappa \star f + b\right) \tag{1.13}$$

is still translation equivariant; it still commutes with the translation operator. We can therefore build arbitrarily long sequences of layers (1.13) that preserve translation equivariance. That being said, there are also other layers that break equivariance. One example is *pooling layers* which are used for downsampling; essentially, throwing away (hopefully redundant) information in order to speed up computations. To give an explicit example, *max pooling layers* split $\mathbb{Z}^2$ into small components and computes the maximum value of data points $f$ in each component, which clearly breaks equivariance (Figure 1.3).

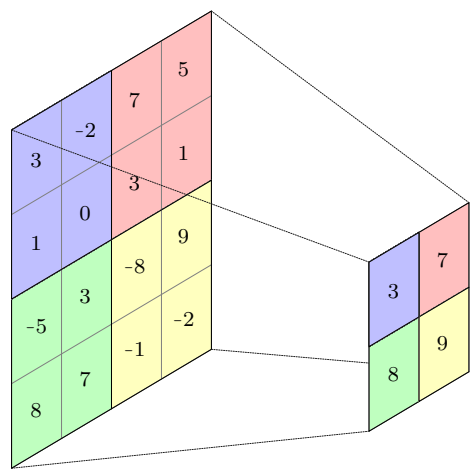



**Figure 1.3:** Max pooling layer.

Translation equivariance is an important property of convolutional layers that facilitates learning, but it is not the only aspect that determines the performance of a CNN. Sometimes, achieving higher performance may require using pooling layers or other layers that break equivariance. It is also common practice to use a multilayer perceptron or some other neural network for the last layers in a CNN, and these networks also break equivariance. However, the general idea

is that if equivariance aids in the extraction of useful features in the first layers, it may not be needed later on.

### 1.1.3 Geometric deep learning

In many deep learning tasks, data is essentially treated as arrays of numbers that have no geometric or similar useful properties. It is common, for instance, to reshape 2D data into 1D arrays by stacking the columns on top of each other, even though this can destroy any 2D geometry that may have been present.

*Geometric deep learning* models instead utilize geometry, when present in data, to improve learning and achieve higher performance. CNNs are perhaps the simplest examples of geometric deep learning models, because convolutional layers make use of the global translation symmetry in $\mathbb{Z}^2$. On the other hand, CNNs still assume that data points are flat 2D or 3D arrays of numbers.

Other types of data have natural curvature, one example being meteorological data on intercontinental regions. Another example is spherical photography, such as the images in Google Street View. Furthermore, aerial photographs may have a flat geometry and can be processed using ordinary CNNs, but there is an orientation ambiguity in aerial photographs that CNNs do not understand. When processing such images, it would be desirable to use convolutional layers that are equivariant not only to translations, but also to planar rotations.

In all of these examples, data points can be interpreted as vector-valued maps $f : \mathcal{M} \to V$ defined on more general manifolds than $\mathcal{M} = \mathbb{Z}^2$. Spherical images are functions $S^2 \to \mathbb{R}^3$ defined on the sphere $S^2$, for example. Geometric deep learning models are intended for these more general types of data. In **Paper I**, we investigate a particular geometric deep learning model that generalizes the translation equivariance in CNNs, to other types of equivariance for data points defined on other globally symmetric spaces $\mathcal{M}$. This model is called *$G$-equivariant convolutional neural networks (GCNNs)*, where $G$ is the symmetry group that characterizes the global symmetry in $\mathcal{M}$. The symmetry group of the sphere $\mathcal{M} = S^2$, for example, is the rotation group $G = SO(3)$.

A detailed discussion on GCNNs and other geometric deep learning models requires the notions of principal bundles and associated vector bundles which, intuitively, are manifolds with certain spaces (fibers) attached to each point. We give an introduction to these notions in Section 1.3 but in preparation, we also need to discuss group representations.

## 1.2 Representation theory

Equivariant neural networks such as GCNNs utilize symmetry, when present in the vector-valued data functions. Mathematicians study abstract symmetry transformations using group theory, and when these symmetry transformations manifest themselves in terms of linear transformations on vector spaces, they fall under the label *representation theory*.

### 1.2.1 Lie groups and Haar measures

Before introducing representations, let us briefly summarize a few relevant notions: Lie groups, Haar measures, and unimodularity. See Lee (2013) for a more comprehensive introduction to Lie groups. Haar measures on locally compact groups, and the concept of unimodularity, is discussed in detail in Deitmar and Echterhoff (2014) and Folland (2016).

**Definition 1.** A *Lie group* is a group $G$ that is also a (real) smooth manifold, such that the following multiplication and inversion maps are smooth:

$$\begin{aligned} &m : G \times G \to G, && m(g,h) = gh, \\ &i : G \to G, && i(g) = g^{-1}. \end{aligned} \tag{1.14}$$

**Remark 1.** Smooth manifolds are assumed Hausdorff and second countable. The same naturally goes for Lie groups, being instances of smooth manifolds.

**Example 1.**

(a) Euclidean space $\mathbb{R}^d$ is an abelian $d$-dimensional Lie group.

(b) Let $G$ be a finite or countable group equipped with the discrete topology. There is a unique smooth structure on $G$ that turns $G$ into a 0-dimensional Lie group. For instance, $\mathbb{Z}^2$ is a Lie group.

(c) The general linear groups $GL(n,\mathbb{R})$ and $GL(n,\mathbb{C})$ are Lie groups for all natural numbers $n \geq 1$. Many more Lie groups are obtained as closed subgroups of these: $O(n)$, $SO(n)$, $U(n)$, $SU(n)$, $SL(n,\mathbb{R})$, $SL(n,\mathbb{C})$, etc.

(d) The Euclidean groups $E(n)$ are Lie groups for all natural numbers $n \geq 1$, as are the Poincaré group, the Lorentz group, and other, similar groups.

■

The smooth manifold structure of a Lie group $G$ implies that each point $g \in G$ has a compact neighbourhood. Hausdorff topological groups satisfying this local compactness property are aptly known as *locally compact groups*, and form the backbone of harmonic analysis.

**Theorem 1** (Folland, 2016, §2.2)**.** *Let $G$ be a locally compact group. There exists a non-zero Radon measure $\mu$ on $G$, called a* (left) Haar measure*, that is left-invariant in the sense that*

$$\mu(gS) = \mu(S), \tag{1.15}$$

*for all measurable sets $S \subseteq G$ and all $g \in G$. The Haar measure $\mu$ on $G$ is unique up to scaling and, if $G$ is compact, the scaling can be chosen such that $\mu(G) = 1$.*

We typically assume that a scale has been chosen and refer to *the* Haar measure on $G$. There is also an equivalent notion of *right* Haar measures, which satisfy $\mu(Sg) = \mu(S)$, but these two notions do not coincide in general.

**Definition 2.** A locally compact group $G$ is *unimodular* if any left Haar measure on $G$ is also a right Haar measure.

Unimodular groups include some of the most commonly studied groups, such as compact groups, abelian groups, finite and discrete groups, the Euclidean groups $E(n)$, and so on.

One may integrate compactly supported continuous functions $f \in C_c(G)$ with respect to the Haar measure $\mu$ on $G$. Left-invariance implies that for all $h \in G$,

$$\int_G f(hg)\,\mathrm{d}\mu(g) = \int_G f(g)\,\mathrm{d}\mu(g). \tag{1.16}$$

Unimodular groups also satisfy the right-invariance and inversion properties

$$\int_G f(gh)\,\mathrm{d}\mu(g) = \int_G f(g)\,\mathrm{d}\mu(g) = \int_G f(g^{-1})\,\mathrm{d}\mu(g). \tag{1.17}$$

A natural question is how integration with respect to the Haar measure relates to integration of differential forms, when $G$ is a Lie group. This question is answered in (Wallach, 2018, §2.5) by constructing left Haar measures on $G$ as top forms $\omega$ induced from alternating top forms on the *Lie algebra* $\mathfrak{g}$, i.e., the space of left-invariant vector fields on $G$. This construction nicely illustrates why different choices of alternating forms yield the same Haar measure up to scaling, and how right Haar measures are obtained from alternating forms on right-invariant vector fields. The moral of the story is that integration in the

Haar sense coincides with integration in the sense of smooth manifolds:

$$\int_G f \, \mathrm{d}\mu = \int_G f\omega, \tag{1.18}$$

assuming compatible scalings of $\mu$ and $\omega$. Let us ignore the integration theoretic technicalities from now on and simply write the integral of $f \in C_c(G)$ as

$$\int_G f(g) \, \mathrm{d}g, \tag{1.19}$$

assuming any choice of left Haar measure on the Lie group $G$.

Before moving on to representations, observe that the assignment

$$\langle f, f' \rangle_2 := \int_G f(g)\overline{f'(g)} \, \mathrm{d}g, \qquad f, f' \in C_c(G), \tag{1.20}$$

defines an inner product on $C_c(G)$. Let $L^2(G)$ be the completion of $C_c(G)$ after performing the usual identification of functions that agree almost everywhere:

$$f = f' \iff \|f - f'\|_2 = 0, \quad f, f' \in L^2(G). \tag{1.21}$$

The Hilbert space $L^2(G)$ is separable when $G$ is second countable. In particular, this holds when $G$ is a Lie group. Much of the analysis in **Paper I** is related to $L^2(G)$ and its use in harmonic analysis - we will say more about this below.

### 1.2.2 Group representations

It is time for us to discuss group representations. These can be motivated from at least two perspectives: Representations make it possible to translate group theoretic problems into functional analysis, which often makes them easier to solve. Conversely, many geometric problems are naturally formulated in terms of linear transformations on vector spaces, and often involve symmetry of some kind. Group theory is the formal study of symmetry, so it is not surprising that we can view such problems from a group theory point of view.

**Definition 3.** Let $G$ be a topological group and let $V$ be a topological vector space. A *(strongly continuous) representation* of $G$ is a group homomorphism

$$\rho : G \to GL(V) \tag{1.22}$$

such that the following map is continuous for all $g \in G$, $v \in V$:

$$G \times V \to V, \qquad (g, v) \mapsto \rho(g)v. \tag{1.23}$$

Representations are denoted $(\rho, V)$ or simply $\rho$. Finally, a representation $(\rho, V)$ is called *unitary* if $V$ is a complex Hilbert space and $\rho(g)$ is unitary for all $g \in G$.

**Remark 2.** We give some examples of representations on vector spaces over $\mathbb{R}$, and representations can be studied for vector spaces over any field $\mathbb{K}$. However, complex representations are especially well-behaved thanks to the fundamental theorem of algebra and other strong results. Introductory texts usually focus on complex representations for this reason, and we make the same choice: Unless otherwise stated, vector spaces $V$ are assumed to be complex.

**Example 2.** Let $G$ be a topological group. For any topological vector space $V$, there is a *trivial representation* that sends each $g \in G$ to the identity operator,

$$\rho : G \to GL(V), \qquad \rho(g) = \mathrm{Id}_V\,. \tag{1.24}$$

The special case $V = \mathbb{C}$ is known as *the* trivial representation. ■

**Example 3.** Consider the rotation group $SO(2)$ and identify each element with its angle of rotation $\theta$. Then the map

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{1.25}$$

becomes a representation, on $\mathbb{R}^2$ as well as on $\mathbb{C}^2$. Indeed, for all angles $\theta, \phi$,

$$\begin{aligned} R(\theta)R(\phi) &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{bmatrix} = R(\theta+\phi) \end{aligned}$$

In particular, $R(\theta)R(-\theta) = R(0) = \mathrm{Id}_{\mathbb{R}^2}$, hence $R(-\theta) = R(\theta)^{-1}$. ■

**Example 4.** Let $S_n$ be the symmetric group whose elements are permutations of $n$ objects, for $n = 1, 2, 3, \ldots$. That is, the elements $g \in S_n$ are bijections

$$g : \{1, 2, 3, \ldots, n\} \to \{1, 2, 3, \ldots, n\}. \tag{1.26}$$

Given a basis $e_1, \ldots, e_n \in \mathbb{C}^n$ and a permutation $g \in S_n$, there exists a unique linear operator $\rho(g)$ that permutes the basis vectors: $\rho(g)e_i = e_{g(i)}$ for each $i = 1, \ldots, n$. The mapping $g \mapsto \rho(g)$ is a representation of $S_n$. ■

The next definition shows how to build representations from simpler ones.

**Definition 4.** The *direct sum* of two $G$-representations $(\rho, V_\rho)$ and $(\sigma, V_\sigma)$ is the representation $(\rho \oplus \sigma, V_\rho \oplus V_\sigma)$ defined by

$$(\rho \oplus \sigma)(g) : v \oplus w \mapsto \rho(g)v \oplus \sigma(g)w. \tag{1.27}$$

Similarly, their *tensor product* is the representation $(\rho \otimes \sigma, V_\rho \otimes V_\sigma)$ defined by

$$(\rho \otimes \sigma)(g) : v \otimes w \mapsto \rho(g)v \otimes \sigma(g)w. \tag{1.28}$$

We can also deconstruct some representations into smaller ones:

**Definition 5.** Let $(\rho, V)$ be a $G$-representation and suppose that $W \subseteq V$ is a linear subspace that is closed under the representation,

$$\rho(g)w \in W, \qquad \text{for all } w \in W, g \in G. \tag{1.29}$$

Then $W$ is an *invariant subspace*, and $(\rho, W)$ is a *subrepresentation* of $(\rho, V)$.

All representations $(\rho, V)$ have two obvious subrepresentations, obtained by letting $W \subseteq V$ be either the full space $W = V$ or the trivial subspace $W = \{0\}$. These subrepresentations are not very interesting and are not considered *proper*, because $W$ is not a proper subspace.

**Definition 6.** A representation is *reducible* if it has a proper subrepresentation. Representations that are not reducible are called *irreducible*.

**Example 5.** A trivial representation $(\mathrm{Id}_V, V)$ is irreducible iff $\dim V = 1$. ■

The next lemma explains why unitary representations are especially interesting: They can be decomposed into direct sums of subrepresentations.

**Lemma 2.** *Let $(\rho, V)$ be a unitary representation and suppose that $W \subseteq V$ is a closed, invariant subspace. Then the orthogonal complement $W^\perp$ is also a closed invariant subspace, and $(\rho, V)$ decomposes into a direct sum representation on $W \oplus W^\perp$.*

*Proof.* Recall the definition of the orthogonal complement,

$$W^\perp = \left\{ w^\perp \in V_\rho \;\middle|\; \langle w, w^\perp \rangle = 0 \text{ for all } w \in W \right\}. \tag{1.30}$$

According to a standard result in functional analysis, if $V$ is a Hilbert space and $W \subseteq V$ is a closed subspace, then $W^\perp$ is a closed subspace and $V = W \oplus W^\perp$. In particular, $W$ and $W^\perp$ are Hilbert spaces in their own right. It is therefore

sufficient to prove that $W^\perp$ is an invariant subspace, i.e., that $\rho(g)w^\perp \in W^\perp$ for each $w^\perp \in W^\perp$ and all $g \in G$. But this follows immediately from the facts that $\rho$ is unitary and that $W$ is an invariant subspace:

$$\langle w, \rho(g)w^\perp \rangle = \langle \rho(g^{-1})w, w^\perp \rangle = 0, \tag{1.31}$$

for all $w \in W$. That is, $\rho(g)w^\perp \in W^\perp$ and the lemma follows. □

It seems likely that unitary representations can be decomposed into direct sums of *irreducible* subrepresentations, if we just keep decomposing a representation into smaller and smaller subrepresentations until these cannot be decomposed any further. This is certainly true for compact groups, which follows from the famous Peter-Weyl theorem (Deitmar and Echterhoff, 2014, Theorem 7.2.4). It is also true for many other groups if we replace direct sums with *direct integrals*. In this sense, irreducible representations are the prime numbers of representation theory, the building blocks used to construct all other representations.

It is sometimes possible to translate one representation into another. An almost trivial example is that the real $SO(2)$-representation (1.25) can be transformed into the following real $SO(2)$-representation,

$$\sigma(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & \\ \sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{1.32}$$

that performs rotations about the $y$-axis in $\mathbb{R}^3$. All we need to do is identify $\mathbb{R}^2$ with the $xz$-plane in $\mathbb{R}^3$ via the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^3$ given by

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{1.33}$$

in the standard bases. If we fix a vector $v \in \mathbb{R}^2$ and a rotation angle $\theta$, it seems likely that the following two procedures give the same result:

1. First use $R(\theta)$ to rotate $v$ and then map the rotated vector to the $xz$-plane.
2. First map $v$ to the $xz$-plane and then use $\sigma(\theta)$ to rotate about the $y$-axis.

Indeed, a short calculation shows that $T \circ R(\theta) = \sigma(\theta) \circ T$. In this example, the relation between $R(\theta)$ and $\sigma(\theta)$ was rather obvious, but this way of relating two representations can be applied much more generally.

**Definition 7.** An *intertwiner*, or *equivariant map*, between two $G$-representations $(\rho, V_\rho)$, $(\sigma, V_\sigma)$ is a bounded linear map $T : V_\rho \to V_\sigma$ satisfying, for all $g \in G$,

$$T \circ \rho(g) = \sigma(g) \circ T. \tag{1.34}$$

We let $\mathrm{Hom}_G(V_\rho, V_\sigma)$ denote the space of all such intertwiners. Moreover, if an intertwiner $T \in \mathrm{Hom}_G(V_\rho, V_\sigma)$ is a (unitary) isomorphism, then $(\rho, V_\rho)$, $(\sigma, V_\sigma)$ are said to be *(unitarily) equivalent*.

**Example 6.** When viewed as a complex representation, the $SO(2)$-representation (1.25) is unitarily equivalent to the direct sum representation $\rho(\theta) = e^{i\theta} \oplus e^{-i\theta}$:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = T \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix} T^*, \tag{1.35}$$

where $T = \frac{1}{\sqrt{2}} \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$. ■

Irreducible representations have strict limitations on the possible intertwiners, as the next lemma shows. This is a famous result known as *Schur's first lemma*.

**Lemma 3** (Deitmar and Echterhoff, 2014, Lemma 6.1.7)**.** *Let $(\rho, V)$ be a unitary representation of a topological group $G$. Then the following are equivalent:*

*(a) $(\rho, V)$ is irreducible.*

*(b) If $T : V \to V$ is an intertwiner, there is a constant $\lambda \in \mathbb{C}$ such that $T = \lambda \,\mathrm{Id}$.*

This lemma is motivated by the following observation: Let $(\rho, V)$ be a unitary representation of $G$ and suppose that $T : V \to V$ is an intertwiner. If $\lambda$ is an eigenvalue of $T$, then the corresponding eigenspace $E_\lambda$ is an invariant subspace of $V$ since, for each $v \in E_\lambda$ and all $g \in G$,

$$T\rho(g)v = \rho(g)Tv = \lambda\rho(g)v, \tag{1.36}$$

hence $\rho(g)v \in E_\lambda$. The unitary representation $(\rho, V)$ is thus reducible if $E_\lambda$ is a proper subspace. Equivalently, if $(\rho, V)$ is irreducible, we must have $E_\lambda = V$ which means that $Tv = \lambda v$ for all $v \in V$ and so $T = \lambda \,\mathrm{Id}$. This is not a complete proof because it assumes the existence of an eigenvalue, but it provides some intuition. The other direction is more straightforward.

The next result, *Schur's second lemma*, is even more famous than the first. It is proven by noting that for any intertwiner $T : V_\rho \to V_\sigma$, its adjoint $T^*$ is also an intertwiner, hence so is their composition $T^*T : V_\rho \to V_\rho$. Schur's first lemma

can thus be applied to the positive semi-definite map $T^*T$, which allows us to extract the relevant information about $T$.

**Corollary 4** (Deitmar and Echterhoff, 2014, Corollary 6.1.9)**.** *Let $(\rho, V_\rho)$, $(\sigma, V_\sigma)$ be irreducible unitary representations and assume that $T : V_\rho \to V_\sigma$ is an intertwiner. Then $T$ is either zero or invertible with continuous inverse. In the latter case there is a scalar $c > 0$ such that $cT$ is unitary. The space $\mathrm{Hom}_G(V_\rho, V_\sigma)$ is zero unless $\rho$ and $\sigma$ are unitarily equivalent, in which case the space has dimension 1.*

We end this part with a discussion on the left and right regular representations of a unimodular Lie group $G$. For reasons that are explained in **Paper I**, these representations are tightly connected both to convolutional layers in GCNNs and to Fourier analysis. This connection lets us apply Fourier theoretic methods when studying GCNNs.

**Example 7.** Let $G$ be a unimodular Lie group. For each $g \in G$, define the map

$$\lambda(g) : L^2(G) \to L^2(G), \qquad (\lambda(g)f)(g') = f(g^{-1}g'), \tag{1.37}$$

for $f \in L^2(G)$ and $g' \in G$. The assignment $g \mapsto \lambda(g)$ is a unitary representation of $G$ called the *left regular representation*. It is unitary thanks to left-invariance of the Haar measure:

$$\begin{aligned}\langle \lambda(g)f, \lambda(g)f' \rangle &= \int_G f(g^{-1}g')\overline{f'(g^{-1}g')}\,\mathrm{d}g' = [g' \mapsto gg'] \\ &= \int_G f(g')\overline{f'(g')}\,\mathrm{d}g' = \langle f, f' \rangle\end{aligned}$$

Similarly, one can define a unitary *right regular representation* $\rho$ by

$$\rho(g) : L^2(G) \to L^2(G), \qquad (\rho(g)f)(g') = f(g'g), \tag{1.38}$$

for $f \in L^2(G)$ and $g, g' \in G$. ■

**Lemma 5.** *The left and right regular representations of a unimodular Lie group $G$ are unitarily equivalent.*

*Proof.* Define an involution operator $T : L^2(G) \to L^2(G)$ by $(Tf)(g) = f(g^{-1})$. This is clear a unitary map that intertwines the left and right representations:

$$\big[T \circ \lambda(g)f\big](g') = f\big((g^{-1}g')^{-1}\big) = f\big(g'^{-1}g\big) = \big[\rho(g) \circ Tf\big](g'), \tag{1.39}$$

for $f \in L^2(G)$ and $g, g' \in G$. □

## 1.3 Fiber bundles

We now turn our attention to fiber bundles, which form the natural setting for geometric data functions on manifolds. Let us introduce the subject with a couple of motivating examples. More detailed expositions of fiber bundles can be found in Kolár et al. (2013) and Lee (2013).

Consider a smooth manifold $\mathcal{M}$ of dimension $d$ and recall that, for each point $x \in \mathcal{M}$, there is an associated tangent space $T_x\mathcal{M}$. A *vector field* on $\mathcal{M}$ assigns a tangent vector $X_x \in T_x\mathcal{M}$ to each point in some open set $U \subseteq \mathcal{M}$ and this assignment should be at least continuous. But the tangent vectors $X_x$ lie in different tangent spaces $T_x\mathcal{M}$ for different points $x$, so we cannot talk about continuous vector fields unless we know how to vary the space $T_x\mathcal{M}$ continuously. The *tangent bundle* $T\mathcal{M}$ solves this problem by defining a suitable topology and smooth structure on the disjoint union of all tangent spaces,

$$T\mathcal{M} := \dot{\bigcup_{x\in\mathcal{M}}} T_x\mathcal{M}. \tag{1.40}$$

The structure on $T\mathcal{M}$ is defined in such a way that the projection $\pi : T\mathcal{M} \to \mathcal{M}$, $X_x \mapsto x$, that sends a tangent vector to the point it is attached to on the manifold, is smooth. Vector fields are then defined as continuous or smooth maps

$$X : U \to T\mathcal{M}, \quad \text{satisfying} \quad \pi \circ X = \mathrm{Id}_{\mathcal{M}}, \tag{1.41}$$

which ensures that $X(x)$ lies in the correct tangent space $T_x\mathcal{M}$ for each $x \in U$. We picture the tangent bundle as a copy of the manifold $\mathcal{M}$ with the tangent space $T_x\mathcal{M}$ attached to each point $x \in \mathcal{M}$. Given any local coordinate chart $(u^1, \ldots, u^d) : U \to \mathbb{R}^d$, for $U \subset \mathcal{M}$, recall that the tangent vectors

$$\left.\frac{\partial}{\partial u^i}\right|_x \in T_x\mathcal{M}, \qquad i = 1, \ldots, d, \qquad x \in U, \tag{1.42}$$

form a basis in $T_x\mathcal{M}$ called a *coordinate basis*. Note that the maps $x \mapsto \left.\frac{\partial}{\partial u^i}\right|_x$ are examples of (smooth, local) vector fields on $U$ for $i = 1, \ldots, d$. The coordinate basis also gives an isomorphism $T_x\mathcal{M} \simeq \mathbb{R}^d$ for each $x \in U$, so the tangent bundle can locally be viewed as the cartesian product $U \times \mathbb{R}^d$. One might expect this idea to extend to the whole bundle, so that $T\mathcal{M}$ would be isomorphic to $\mathcal{M} \times \mathbb{R}^d$. If so, we would call $T\mathcal{M}$ a *trivlal* bundle. However, the lack of a global coordinate chart (in general) prevents this idea from coming to fruition; the tangent bundle generally has a more intricate geometric structure and is not isomorphic to $\mathcal{M} \times \mathbb{R}^d$.

Lie groups $G$ have a similar geometric structure: If $K \leq G$ is a closed subgroup, then $G$ is the disjoint union of cosets

$$G = \dot{\bigcup_{g \in G}} gK, \tag{1.43}$$

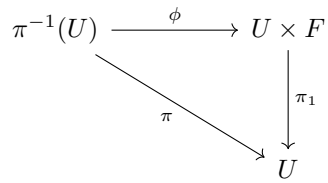
and we may think of $G$ as the quotient manifold $G/K$ with the cosets $gK \simeq K$ attached to each point. The quotient map $q : G \to G/K$ then sends each group element $g \in G$ to the point $q(g) \in G/K$ it is attached to, just like the projection $\pi : T\mathcal{M} \to \mathcal{M}$ of the tangent bundle. We may also define analogues of vector fields as continuous maps $s : U \to G$ satisfying $q \circ s = \mathrm{Id}_{G/K}$, where $U \subseteq G/K$ is an open set. Further, $G$ can locally be viewed as a cartesian product $U \times K$ but again, this does not hold globally: $G$ is generally not isomorphic to $G/K \times K$.

In both of these examples, we attached a *fiber* $(T_x\mathcal{M}; gK)$ to each point on a *base space* $(\mathcal{M}; G/K)$, thus obtaining a larger *total space* $(T\mathcal{M}; G)$. A *projection* $(\pi; q)$ connected the total space and the base space, sending each element in the total space to the point in the base space where it is attached. Moreover, the total space could locally be viewed as a cartesian product $(U \times \mathbb{R}^d; U \times K)$ involving a *characteristic fiber* $(\mathbb{R}^d; K)$, but this local perspective did not extend globally. Together, these properties define a *fiber bundle*.

**Definition 8.** A *fiber bundle* is a structure $(E, \pi, X, F)$, where $E$, $X$, and $F$ are topological spaces and $\pi : E \to X$ is a surjective continuous map satisfying the following: For each $x \in X$, there exists a neighbourhood $U \subseteq X$ around $x$ and a homeomorphism

$$\phi : \pi^{-1}(U) \to U \times F, \tag{1.44}$$

called a *local trivialization*, such that the following diagram commutes:




Here, $\pi_1 : U \times F \to U$ is the projection onto the first coordinate.

The topological spaces $E$, $X$, and $F$ are respectively called the *total space*, the *base space*, and the *characteristic fiber* of the bundle. Furthermore, we say that $(E, \pi, X, F)$ is a *smooth* fiber bundle if $E$, $X$, and $F$ are smooth manifolds, $\pi$ is a smooth map, and the local trivializations can be chosen to be diffeomorphisms.

**Remark 3.** It is common to denote fiber bundles by their projection $\pi : E \to X$, or simply by the total space $E$, leaving the other ingredients implicit.

There are many different kinds of fiber bundles, depending on the characteristic fiber $F$ and its properties. When $F$ is a vector space, for example, the bundle could very well be a *vector bundle*.

**Definition 9.** Let $\pi : E \to X$ be a fiber bundle whose fibers $E_x = \pi^{-1}(\{x\})$ have the structure of a finite-dimensional vector space. Assume that, for any local trivialization $\phi : \pi^{-1}(U) \to U \times F$ and each $x \in U$, the map

$$F \to E_x, \qquad v \mapsto \phi^{-1}(x, v), \tag{1.45}$$

is an isomorphism of vector spaces. Then $E$ is called a *vector bundle*.

**Example 8.** The tangent bundle is a smooth vector bundle: Its fibers $\pi^{-1}(\{x\}) = T_x\mathcal{M}$ are $d$-dimensional vector spaces and the local trivializations are obtained from local coordinate charts, hence (1.45) becomes

$$\phi : \mathbb{R}^d \to T_x\mathcal{M}, \qquad \left(v^1, \dots, v^d\right) \mapsto \sum_{i=1}^{d} v^i \frac{\partial}{\partial u^i}\bigg|_x, \tag{1.46}$$

which is an isomorphism of vector spaces. ■

**Definition 10.** Let $\pi : E \to X$ and $\pi' : E' \to X'$ be fiber bundles. A continuous map $\varphi : E \to E'$ is called a *bundle map (bundle morphism)* if there is a continuous map $f : X \to X'$ such that the following diagram commutes:

$$\begin{array}{ccc} E & \xrightarrow{\varphi} & E' \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi'} \\ X & \xrightarrow{f} & X' \end{array}$$

If both $\varphi$ and $f$ are homeomorphisms, the bundle map is called an *isomorphism of bundles*, or a *bundle isomorphim*. Smooth bundle maps are defined analogously.

**Remark 4.** The notion of *isomorphism* introduced in Definition 10 depends on the type of bundle. An *isomorphism of smooth vector bundles*, for example, must be smooth as well as linear on each fiber.

We mentioned earlier that the tangent bundle $T\mathcal{M}$ can be isomorpic to $\mathcal{M} \times \mathbb{R}^d$, even though this is rarely the case. Cartesian products such as $\mathcal{M} \times F$ trivially

satisfy the definition of a fiber bundle, because we can define the projection

$$\pi : \mathcal{M} \times F \to \mathcal{M}, \qquad \pi(x, f) = x, \tag{1.47}$$

by projection onto the first component. Also, the identity map on $\mathcal{M} \times F$ is a local trivialization.

**Definition 11.** A fiber bundle $\pi : E \to \mathcal{M}$ is *trivial* if it is isomorphic to $\mathcal{M} \times F$.

The geometric deep learning models that we consider in **Paper I** assume that data points are collections of vectors attached at different points on a manifold, and may therefore have a geometric structure. The following definition makes this idea precise.

**Definition 12.** Let $\pi : E \to \mathcal{M}$ be a fiber bundle and let $U \subseteq \mathcal{M}$ be an open set. A *(local) section* is a continuous map

$$s : U \to E, \qquad \text{satisfying} \qquad \pi \circ s = \mathrm{Id}_{\mathcal{M}}\,. \tag{1.48}$$

That is, sections map each point $x \in U$ to an element of its fiber $E_x$.

Data points are defined in **Paper I** as (compactly supported) sections of vector bundles. In particular, data points include vector fields, which are sections of the tangent bundle $T\mathcal{M}$. Moreover, we discussed earlier that **Paper I** analyzes GCNNs, which generalize the equivariance properties of convolutional layers in CNNs to data points defined on other globally symmetric spaces $\mathcal{M}$. This is indeed the main focus in **Paper I**, but it also discusses *local* symmetries. In a sense, local symmetries describe certain redundancies or degrees of freedom that exist in a given situation.

Perhaps the best example of a local symmetry is that we can choose a basis in any tangent space $T_x\mathcal{M}$, but we cannot choose a basis for *all* tangent spaces in a continuous manner. Not unless the tangent bundle $T\mathcal{M}$ is trivial. Naively, one might consider defining a geometric deep learning model for vector fields by choosing bases in all tangent spaces, and then working exclusively in terms of components. This would turn vector fields into arrays of numbers and we could probably use standard neural networks to process these. However, it is simply not possible to make a continuous choice of basis for each tangent space, we are forced to work locally on patches. But this introduces a problem: What if two patches overlap and we use two different bases for the same tangent space $T_x\mathcal{M}$ in the two patches? Most neural networks do not understand that two arrays can represent the same vector in different bases, so numerical results obtained in one basis may be incompatible with those obtained in another basis. Moreover, this problem is fundamentally unavoidable. The local symmetry in this scenario is the choice of bases in tangent spaces.

The study of local symmetry falls under the name *gauge theory*, which originates in theoretical physics. It uses the following notions of principal bundles and associated bundles, and we bring these notions up since they are equally useful when studying the globally symmetric setting of GCNNs.

**Definition 13.** Let $K$ be a topological group. A fiber bundle $\pi : P \to \mathcal{M}$ is called a *principal $K$-bundle* with *structure group $K$* if there is a free, continuous right $K$-action

$$P \times K \to P, \qquad (p, k) \mapsto p \triangleleft k, \tag{1.49}$$

with the following properties for each $x \in \mathcal{M}$.

(i) Let $P_x = \pi^{-1}(\{x\})$ be the fiber at $x$. Then

$$p \in P_x,\ k \in K \quad \Rightarrow \quad p \triangleleft k \in P_x. \tag{1.50}$$

That is, the $K$-action preserves fibers.

(ii) For each $p \in P_x$, the mapping $k \mapsto p \triangleleft k$ is a homeomorphism $K \to P_x$.

A principal $K$-bundle $\pi : P \to \mathcal{M}$ is *smooth* if it is a smooth fiber bundle, $K$ is a Lie group, the action (1.49) is smooth, and the maps (1.50) are diffeomorphisms.

**Remark 5.** Sections of principal bundles are called *gauges*.

GCNNs are concerned with the following smooth principal bundle, which we touched upon in the introduction to this part.

**Proposition 6.** *Let $G$ be a Lie group and let $K \leq G$ be a closed subgroup. Then the quotient map*

$$q : G \to G/K, \qquad g \mapsto gK, \tag{1.51}$$

*defines a smooth principal $K$-bundle over $\mathcal{M} = G/K$.*

*Proof.* If we remove the word "smooth", the proposition actually holds for any topological group $G$ such that the closed subgroup $K \leq G$ admits a gauge, and we formulate the proof for such groups. Lie groups admit smooth gauges and $G$ will thus be a smooth principal $K$-bundle (Steenrod, 1960, p. 31-33).

Assume the existence of a (smooth) gauge $\omega : U \to G$ for an open set $U \subset G/K$, and use $\omega$ to define a map $\phi : q^{-1}(U) \to U \times K$ by

$$\phi(g) = \big(q(g), \omega(q(g))^{-1} g\big) = (gK, k^{-1}), \tag{1.52}$$

where $k = g^{-1}\omega(q(g))$. Then $\phi$ is continuous (smooth), and so is its inverse

$$\phi^{-1}(gK, k) = \omega(gK)k, \tag{1.53}$$

hence $\phi$ is a homeomorphism (diffeomorphism). Moreover, it is evident from (1.52) that $q = \pi_1 \circ \phi$, and so $\phi$ is a local trivialization around any point $gK \in U$. As for points $gK \notin U$, fix an arbitrary $\tilde{g} \in G$ and define the map

$$\omega_{\tilde{g}} : \tilde{g}U \to G, \qquad \omega_{\tilde{g}}(gK) = \tilde{g}\omega(gK). \tag{1.54}$$

This is a continuous (smooth) gauge, and induces a local trivialization $\phi_{\tilde{g}}$ in the same manner as above. Because any point $gK \in G/K$ lies in some $\tilde{g}U$ and thus admits a local trivialization $\phi_{\tilde{g}}$, we find that $q : G \to G/K$ is a (smooth) fiber bundle. Finally, observe that right-multiplication $G \times K \to G, (g, k) \mapsto gk$ is a continuous (smooth) right-action that clearly satisfies both of the conditions in Definition 13, so we conclude that $G$ is a (smooth) principal $K$-bundle. □

**Definition 14.** Consider a principal $K$-bundle $\pi : P \to \mathcal{M}$ and suppose that $\rho : K \to GL(V_\rho)$ is a finite-dimensional representation of the structure group. Now define the following equivalence relation $\sim$ on the product $P \times V_\rho$,

$$(p, v) \sim (p \triangleleft k, \rho(k)^{-1}v), \qquad p \in P, v \in V_\rho, k \in K. \tag{1.55}$$

Then the quotient space $P \times_\rho V_\rho = (P \times V_\rho)/\sim$, that is,

$$P \times_\rho V_\rho = \{\text{equivalence classes } [p, v] \mid p \in P, v \in V_\rho\}, \tag{1.56}$$

is called an *associated bundle*.

**Remark 6.** Associated bundles are most often called associated *vector* bundles, since they are special cases of a more general definition. However, we only use associated vector bundles here and in **Paper I**, so we use the shorter name.

In **Paper I**, we give an example that shows explicitly how the tangent bundle $T\mathcal{M}$ is an associated bundle. More importantly, we demonstrate that GCNNs are naturally formulated in terms of *homogeneous vector bundles*, which are also associated bundles. In other words, this concept is very useful.

Of course, one cannot simply claim that associated bundles $P \times_\rho V_\rho$ are bundles by virtue of their name, it must be proven. One proof uses a gauge $\omega : U \to P$ to pick representatives $(\omega(x), v) \in P \times V_\rho$ of equivalence classes $[\omega(x), v] \in P \times_\rho V_\rho$, and these representatives are then used to construct the local trivializations. A complete proof can be found in (Kolár et al., 2013, §10.7)

**Proposition 7.** *Let $P$ be a (smooth) principal $K$-bundle and suppose that $(\rho, V_\rho)$ is a finite-dimensional $K$-representation. Then $P \times_\rho V_\rho$ is a (smooth) vector bundle.*

We end with a result that explains the aforementioned problem with choosing bases in tangent spaces. The process of continously choosing such bases defines a gauge in a principal bundle called a *frame bundle* $F\mathcal{M}$ and for most manifolds, the frame bundle is nontrivial.

**Proposition 8.** *A principal $K$-bundle $P$ is trivial iff it admits a global gauge.*

*Proof.* First suppose that $P$ is trivial, and let $\varphi : \mathcal{M} \times K \to P$ be an isomorphism of bundles. Then the mapping

$$\omega : \mathcal{M} \to P, \qquad \omega(x) = \varphi(x, k), \tag{1.57}$$

is a continuous, global gauge for any fixed $k \in K$. For the other direction, suppose $\omega : \mathcal{M} \to P$ is a global gauge and define the mapping

$$\phi : \mathcal{M} \times K \to P, \qquad (x, k) \mapsto \omega(x) \triangleleft k, \tag{1.58}$$

which is continuous since both $\omega$ and the $K$-action are continuous. If we now fix $x \in \mathcal{M}$, then the restriction $\phi_x : K \to P_x$, $k \mapsto \phi(x, k)$ is a homeomorphism by Definition 13(ii), hence (1.58) has a continuous inverse

$$\phi^{-1} : P \to \mathcal{M} \times K, \qquad p \mapsto \big(\pi(p), \phi^{-1}_{\pi(p)}(p)\big). \tag{1.59}$$

Furthermore, (1.58) preserves basepoints and satisfies

$$\phi(x, kk') = \omega(x) \triangleleft k \triangleleft k' = \phi(x, k) \triangleleft k', \tag{1.60}$$

and is therefore an isomorphism of principal bundles. That is, $P$ is trivial. □

# 2 Summary of results

In this chapter, we provide a summary of the paper included in this thesis. The overall aims, findings, and conclusions from the paper are presented to further make it easier to understand the paper's contribution to the research field.

In **Paper I**, *Homogeneous vector bundles and G-equivariant convolutional neural networks*, we conduct a mathematical analysis of *G-equivariant convolutional neural networks (GCNNs)*. These networks are aimed at deep learning tasks in which the data is defined on a globally symmetric space $\mathcal{M}$ with global symmetry group $G$, for example spherical data ($\mathcal{M} = S^2, G = SO(3)$). Incorporating the global symmetry of $\mathcal{M}$ into the network structure, through *G-equivariant layers*, reduces the amount of information that the network has to learn.

Local symmetry is another type of symmetry that describes redundancies in mathematical models; degrees of freedom which give rise to degeneracies or other problems and cannot be ignored, even if they have no effect on the model predictions. One such example is the choice of basis in a vector space (or more generally, the choice of local frame in a tangent bundle). Such choices might be necessary in order to perform computations but should have no effect on what the model predicts. Neural networks are not designed with local symmetry in mind, they can be highly dependent on how the data is represented. *Gauge equivariant neural networks* were introduced to counteract this problem. These networks are especially useful when the data is defined on a curved manifold, where one is forced to work locally in patches but want to avoid inconsistencies on overlaps.

GCNNs and gauge equivariant neural networks are both based on the gauge theoretic notions of principal bundles and associated vector bundles. However, these two network types have only been studied separately from each other. We combine the two network types into a common framework, highlighting their many similarities. In fact, their only significant difference is the equivariance properties imposed on layers.

We also show that GCNNs arise naturally from homogeneous vector bundles. These bundles are closely related to the aforementioned equivariance framework, but have not been previously discussed in this context. In our opinion, GCNNs should be formally defined in terms of homogeneous vector bundles, as these bundles perfectly encapsulate the idea of globally symmetric spaces and $G$-equivariant transformations.

Implementations of GCNNs are typically based on *convolutional layers*. These are special cases of the $G$-equivariant layers mentioned earlier, but have a more explicit structure as integral operators. We give an example of a $G$-equivariant layer that is not a convolutional layer. There have been previous investigations into the question of when a $G$-equivariant layer is, in fact, a convolutional layer. These investigations have either made strong assumptions on layers or have been restricted to compact groups $G$. We answer this question for more general, unimodular Lie groups $G$, using reproducing kernel Hilbert spaces (RKHS) as well as the related concept of bandwidth in Fourier analysis. To our knowledge, this is the first time RKHS have been applied to equivariant neural networks, even though RKHS are used in other areas of deep learning.

# Bibliography

Deitmar, A. and Echterhoff, S. (2014). *Principles of harmonic analysis*. Springer International Publishing.

Folland, G. B. (2016). *A course in abstract harmonic analysis*. CRC press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.

Kolár, I., Michor, P. W., and Slovák, J. (2013). *Natural operations in differential geometry*. Springer Science & Business Media.

Lee, J. M. (2013). *Introduction to smooth manifolds*. Springer.

Steenrod, N. (1960). *The topology of fibre bundles*. Princeton University Press.

Wallach, N. R. (2018). *Harmonic analysis on homogeneous spaces*. M. Dekker.

# Paper I

# HOMOGENEOUS VECTOR BUNDLES AND $G$-EQUIVARIANT CONVOLUTIONAL NEURAL NETWORKS

JIMMY ARONSSON

ABSTRACT. $G$-equivariant convolutional neural networks (GCNNs) is a geometric deep learning model for data defined on a homogeneous $G$-space $\mathcal{M}$. GCNNs are designed to respect the global symmetry in $\mathcal{M}$, thereby facilitating learning. In this paper, we analyze GCNNs on homogeneous spaces $\mathcal{M} = G/K$ in the case of unimodular Lie groups $G$ and compact subgroups $K \leq G$. We demonstrate that homogeneous vector bundles is the natural setting for GCNNs. We also use reproducing kernel Hilbert spaces to obtain a precise criterion for expressing $G$-equivariant layers as convolutional layers. This criterion is then rephrased as a bandwidth criterion, leading to even stronger results for some groups.

## CONTENTS

1. Introduction 1
2. Foundations of equivariant neural networks 3
2.1. Convolutional neural networks 3
2.2. Gauge theory and the equivariant framework 4
3. $G$-equivariant convolutional neural networks 9
3.1. Homogeneous spaces 9
3.2. Homogeneous vector bundles 9
3.3. Induced representations 10
3.4. $G$-equivariant and convolutional layers 12
3.5. RKHS and bandlimited functions 17
4. Discussion 20
References 21

## 1. INTRODUCTION

Developments in deep learning have increased dramatically in recent years. Even though multilayer perceptrons [2] and other general-architecture models work well for some tasks, achieving higher levels of performance often requires models that are more tailored to each application, and which incorporate some level of understanding of the data. *Geometric deep learning* [5, 6, 7, 13, 38] is the approach of using inherent geometric structure in data, and symmetry derived from geometry, to improve deep learning models.

*Convolutional neural networks (CNNs)* are among the simplest and most broadly applicable general-architecture models. They have been successfully applied to image classification and segmentation [41, 52, 53], text summarization [42], pose estimation [37], sign language recognition [27], and many other tasks. One reason why CNNs are so useful is that *convolutional layers*, the basic building blocks of CNNs, commute

*Date*: May 12, 2021.

with the translation operator in $\mathbb{Z}^2$; convolutional layers are *translation equivariant*. In image classification tasks, for instance, $\mathbb{Z}^2$ represents the underlying pixel lattice, and translation equivariance helps CNNs identify objects in images regardless of their exact pixel coordinates. As convolutional layers respect the global translation symmetry in $\mathbb{Z}^2$, CNNs are examples of geometric deep learning models.

*G-equivariant convolutional neural networks (GCNNs)* [10, 12] are generalizations of CNNs to data points defined on homogeneous $G$-spaces $\mathcal{M}$. Convolutional layers that commute with the action $G \times \mathcal{M} \to \mathcal{M}$ of the global symmetry group $G$, remove the need for GCNNs to learn about the global symmetry. It is already built into the network. This enables GCNNs to focus on learning other relevant features in data, potentially improving performance. One example is the detection of tumors in digital pathology. Images of tumors can have any orientation, and GCNNs with both translation and rotation equivariant layers have higher accuracy than ordinary CNNs [46]. Rotation equivariance is also highly useful in 3D inference problems [49], in point cloud recognition [35], and in other tasks.

*Gauge equivariant* neural networks [9, 13, 18, 36] are instead designed to respect *local* symmetries. For example, computations involving vector fields - in meteorology or other areas - require vectors to be expressed in components. This requires a frame; a smooth assignment of a basis to each tangent space. However, the sphere and other non-parallelizable manifolds do not admit a global frame, so the computations must be performed locally, using different local frames for different regions on the manifold. It is then important that any numerical results obtained in one frame are compatible with those obtained in any other frame on overlapping regions. In other words, the computations should be equivariant with respect to the choice of local frame, which is a viewed as a gauge degree of freedom; a local symmetry. Gauge equivariant neural networks have also been introduced for problems exhibiting other local symmetries, primarily in lattice gauge theory.

In this paper, we study the mathematical foundations of GCNNs and characterize convolutional layers in terms of more abstract layers. Our contributions are threefold:

- We analyze a general framework that include both gauge equivariant neural networks and GCNNs, that only differ in whether layers respect a local gauge symmetry or a global translation symmetry. Moreover, we show that GCNNs are naturally expressed in terms of homogeneous vector bundles.
- In general, not all $G$-equivariant layers can be written as convolutional layers. We investigate the relation between these types of layers for all homogeneous spaces $\mathcal{M} = G/K$ when $G$ is a unimodular Lie group and $K \leq G$ is a compact subgroup. As a result of this investigation, we find a criterion for expressing $G$-equivariant layers as convolutional layers (Theorem 14).
- We highlight the close relationship between convolutional layers in GCNNs, reproducing kernel Hilbert spaces (RKHS), and bandwidth. We reformulate the criterion in Theorem 14 as a bandwidth criterion and prove that, when $G$ is discrete abelian or finite,[1] all $G$-equivariant layers are indeed convolutional layers (Corollaries 19-21).

This work was inspired by a number of papers [9, 10, 11, 12, 13, 48]. The theoretical papers [9, 12] have been of particular importance, as our work grew from a desire to understand the mathematics of equivariant neural networks in even greater detail.

In the case of compact groups $G$, the Peter-Weyl theorem and other powerful tools have allowed researchers to study GCNNs using harmonic analysis. Among the most

[1]Discrete Lie groups are countable, in this paper, as we assume smooth manifolds to be Hausdorff and second-countable.

well-known results in this direction is Theorem 1 in [28], which uses Fourier analysis on $G$ to establish that the layers in a $G$-equivariant feed-forward neural network must be generalized convolutional layers, when $G$ is compact. This result is similar to our second contribution above and we discuss the distinction in Section 3.4. Others have used the well-known representation theory of the compact group $G = SO(3)$ to study rotation equivariant GCNNs for spherical data [15, 16, 17].

The paper is structured as follows. We summarize the relevant machine learning background in Section 2.1, and discuss a framework for equivariant neural networks in Section 2.2. In Section 3, we restrict attention to homogeneous spaces $G/K$ where $G$ is a unimodular Lie group and $K \leq G$ is a compact subgroup. Section 3.2 explains the relation between GCNNs, homogeneous vector bundles, and induced representations. This relation is used to motivate the definition of $G$-equivariant layers in Section 3.4, where we also discuss convolutional layers and prove the aforementioned Theorem 14; this result characterizes when a $G$-equivariant layer is a convolutional layer, in terms of RKHS. Section 3.5 then relates RKHS to bandwidth, leading to a reformulation of Theorem 14 (Corollary 19) as well as a few stronger results. Finally, in Section 4, we summarize our work and end with a discussion.

## 2. Foundations of equivariant neural networks

In this section, we give an introduction to convolutional neural networks (CNNs) and discuss a simple framework for equivariant neural networks.

### 2.1. Convolutional neural networks.

CNNs were first introduced in 1979 under the name of Neocognitrons, and were used to study visual pattern recognition [21]. In the 1990s, CNNs were successfully applied to problems such as automatic recognition of handwritten digits [33] and face recognition [32]. However, it was arguably not until 2012, when the GPU-based AlexNet CNN outperformed all competition on the ImageNet Large Scale Visual Recognition Challenge [30], that CNNs and other neural networks truly caught the public eye. Industrial work and academic research on deep learning has since soared, and current state-of-the-art deep learning architectures are significantly more powerful and more complex than AlexNet. Yet, convolutional layers remain important components.

In this introduction, we focus on data that can be represented by finitely supported functions

$$f : \mathbb{Z}^2 \to \mathbb{R}^m. \tag{2.1}$$

Digital images, for example, are of this form since each pixel $x \in \mathbb{Z}^2$ is associated with a color array $f(x) \in \mathbb{R}^m$, and finite support is analogous to finite image resolution. Note that $m = 1$ corresponds to grayscale images and $m = 3$ to RGB images, but we allow any number of channels $m$. In general, any data represented by a finite 2D ($m = 1$) or 3D ($m > 1$) array with real-valued entries is of the form (2.1).

*Convolutional layers* act on data points (2.1) by[2]

$$[\kappa \star f](x) = \sum_{y \in \mathbb{Z}^2} \kappa(y - x) f(y), \tag{2.2}$$

given a matrix-valued kernel $\kappa : \mathbb{Z}^2 \to \mathrm{Hom}(\mathbb{R}^m, \mathbb{R}^n)$ for some $n \in \mathbb{N}$. The kernel is also finitely supported in practice, so the maps $\kappa \star f : \mathbb{Z}^2 \to \mathbb{R}^n$ are themselves data points (2.1) with $n$ channels. Broadly speaking, CNNs consist of convolutional layers (2.2) combined with other transformations, such as non-linear activation functions and batch normalization layers. We are mainly interested in convolutional layers,

[2]The name *convolutional layer* is used even though (2.4) more closely resembles a cross-correlation. It can be expressed as a convolution if we replace the kernel with its involution $\kappa^*(y) = \kappa(-y)$.

so we do not go into detail about non-linear activation functions or other types of layers. For more extensive descriptions of CNNs, see [1, 24, 51].

In image classification tasks, for instance, CNNs categorize digital images into a predefined number $k$ of distinct classes, based on what the images depict. The CNN maps each digital image $f : \mathbb{Z}^2 \to \mathbb{R}^m$ to a probability vector in $\mathbb{R}^k$ estimating the probability that $f$ belongs to any given class. During training, this probability vector is compared to the correct answer (which is known) and the discrepancy is computed using a *loss* norm or distance function. A gradient descent-based algorithm minimizes the loss function, thereby *learning* the kernel matrix elements and any other trainable network parameters. The result of this training procedure is a CNN that accurately classifies images in the training data set. Finally, the predictive power of the CNN is evaluated by using it to classify images from a test data set; images that were not used during training and which the CNN has not encountered before.

CNNs perform very well on image classification and similar machine learning tasks, and are important parts of many state-of-the-art network architectures on such tasks [4, 25, 44, 50]. One reason for their success is *translation equivariance*: Convolutional layers (2.2) commute with the translation operator in the image plane,

$$L_x : \mathbb{Z}^2 \to \mathbb{Z}^2, \qquad L_x(y) = y + x, \qquad x \in \mathbb{Z}^2. \tag{2.3}$$

Translation equivariance makes CNNs agnostic to the specific locations of individual pixels, while still taking into account the relative positions of different pixels; images are more easily classified based on relevant features of their subjects, and not based on technical artifacts such as specific pixel coordinates. This observation motivates the introduction of more general convolutional layers that act equivariantly on data points $f : \mathcal{M} \to V$, where the domain $\mathcal{M}$ is homogeneous with respect to a locally compact group $G$ [10, 12]. Given finite-dimensional vector spaces $V, W$, convolutional layers are defined as certain vector-valued integrals[3]

$$\kappa \star f : \mathcal{M} \to W, \qquad (\kappa \star f)(g) = \int_G \kappa(g^{-1}g')f(g') \, \mathrm{d}g', \tag{2.4}$$

with operator-valued kernels $\kappa : G \to \mathrm{Hom}(V, W)$.

**Remark 1.** In (2.4), we integrate with respect to a Haar measure on the unimodular Lie group $G$.

Broadly speaking, $G$-equivariant convolutional neural networks (GCNNs) consist of sequences of convolutional layers (2.4) mixed with non-linear activation functions, and possibly other layers that are equivariant with respect to the global symmetry. This characterization is intentionally vague as we want to avoid making unnecessarily restrictive assumptions on the layers. For this reason, we will not study GCNNs from a holistic perspective, as a sequence of multiple layers, but instead focus on individual layers. We give a formal definition of abstract, $G$-equivariant layers in Definition 9, before defining the more specific convolutional layers in Definition 10.

2.2. **Gauge theory and the equivariant framework.** Before going into detail about GCNNs in Section 3, let us describe a mathematical framework for equivariant neural networks. The framework is based on gauge theoretic concepts but is equally suitable for GCNNs. Both gauge equivariant neural networks and GCNNs will thus be described by this framework, their main difference being the specific equivariance properties imposed on layers.

[3]For a summary on vector-valued integration on locally compact groups, see [19, Appendix 4].

**Remark 2.** This framework is already being used in GCNNs and gauge equivariant neural networks separately [9, 12]. We are simply presenting the unified theory that includes both types of equivariance as separate cases.

Gauge theory originated in physics as a way to model local symmetry. In quantum electrodynamics (QED), for example, the electron wave function can be locally phase shifted, $\psi \mapsto e^{i\alpha}\psi$, with no physically observable consequence, and so QED is said to possess a $U(1)$ gauge symmetry. Mathematicians have later adopted gauge theory in order to study other types of local symmetries. The introduction of gauge equivariant deep learning models has been suggested by deep learning practitioners and physicists alike. For example, [9] investigates the structure of gauge equivariant layers used for vector fields, tensor fields, and more general fields. Physicists have introduced gauge equivariant neural networks for applications in, e.g., lattice gauge theory [3, 18, 36].

We assume some familiarity with fiber bundles,[4] but we still present a few relevant definitions and examples.

**Definition 1.** Let $K$ be a Lie group. A smooth fiber bundle $\pi : P \to \mathcal{M}$ is called a *principal $K$-bundle* with *structure group* $K$ if there is a free, smooth right $K$-action

$$P \times K \to P, \qquad (p, k) \mapsto p \triangleleft k, \tag{2.5}$$

with the following properties for each $x \in \mathcal{M}$.

(i) Let $P_x = \pi^{-1}(\{x\})$ be the fiber at $x$. Then

$$p \in P_x,\ k \in K \quad \Rightarrow \quad p \triangleleft k \in P_x. \tag{2.6}$$

That is, the $K$-action preserves fibers.

(ii) For each $p \in P_x$, the mapping $k \mapsto p \triangleleft k$ is a diffeomorphism $K \to P_x$.

Principal bundles are natural tools for understanding local symmetries, i.e., gauge degrees of freedom. In theoretical physics, gauge degrees of freedom are redundancies in the mathematical theory with no physical relevance. This is both a blessing and a curse: Solving the Yang-Mills equations of motion as an initial value problem, for example, is an underdetermined problem that cannot be solved without taking the gauge degrees of freedom into account; without choosing a *gauge* [39]. This is similar to our example in the introduction, that computations involving vector fields may require a choice of basis in each tangent space, even if this choice is irrelevant for the underlying application. On the other hand, problems may also become easier to solve by choosing a gauge with some finesse.[5]

**Definition 2.** Let $\pi : P \to \mathcal{M}$ be a principal $K$-bundle and assume $U \subseteq \mathcal{M}$ is open.

(i) A *gauge* is a local section $\omega : U \to P$.

(ii) A *gauge transformation* is an automorphism $\chi : P \to P$ that is equivariant,

$$\chi(p \triangleleft k) = \chi(p) \triangleleft k, \qquad p \in P, k \in K, \tag{2.7}$$

and which preserves fibers: $\pi \circ \chi = \pi$.

We will go into more detail about the vector field example in Example 1. However, we first need to define associated bundles. To this end, let $\pi : P \to \mathcal{M}$ be a principal $K$-bundle and let $\rho : K \to GL(V_\rho)$ be a finite-dimensional representation. Define an equivalence relation $\sim$ on $P \times V_\rho$ by

$$(p, v) \sim (p \triangleleft k, \rho(k)^{-1}v), \qquad p \in P, v \in V_\rho, k \in K. \tag{2.8}$$

[4] Introductions to fiber bundles can be found in [26, 34, 40].

[5] See, for example, the temporal gauge in lattice gauge theory [22, §3.3.2].

Let $P \times_\rho V_\rho = (P \times V_\rho)/\sim$ denote the quotient space, whose elements are equivalence classes

$$[p, v] = [p \triangleleft k, \rho(k)^{-1} v], \qquad p \in P, v \in V_\rho, k \in K, \tag{2.9}$$

and consider the projection $\pi_\rho : P \times_\rho V_\rho \to \mathcal{M}$ defined by $\pi_\rho([p, v]) = \pi(p)$. Observe that each fiber $\pi_\rho^{-1}(\{x\})$ has a natural vector space structure such that the mapping

$$V_\rho \to \pi_\rho^{-1}(\{x\}), \qquad v \mapsto [p, v], \tag{2.10}$$

is a linear isomorphism for each fixed $p \in P_x$.

**Lemma 1** ([26, §10.7])**.** *The* associated bundle $\pi_\rho : P \times_\rho V_\rho \to \mathcal{M}$ *is a smooth vector bundle.*

**Example 1.** Let $d = \dim \mathcal{M}$ and consider a coordinate chart $\left(u^1, \ldots, u^d\right) : U \to \mathbb{R}^d$, for $U \subseteq \mathcal{M}$. Recall that for each $x \in U$, the induced coordinate basis in $T_x\mathcal{M}$,

$$\omega(x) := \left( \frac{\partial}{\partial u^1}\Big|_x, \ldots, \frac{\partial}{\partial u^d}\Big|_x \right), \tag{2.11}$$

lets us express tangent vectors $X_x \in T_x\mathcal{M}$ in components $X_x^1, \ldots, X_x^d \in \mathbb{R}$. Moreover, (2.11) defines a local frame $\omega : U \to F\mathcal{M}$ that sends each point $x \in U$ to its coordinate basis in $T_x\mathcal{M}$. Local frames are sections of the frame bundle $F\mathcal{M}$, which is a principal $GL(d, \mathbb{R})$-bundle, hence local frames are examples of gauges.

Fix $x \in U$ and expand $X_x \in T_x\mathcal{M}$ in the coordinate basis:

$$X_x = \sum_{i=1}^{d} X_x^i \frac{\partial}{\partial u^i}\Big|_x. \tag{2.12}$$

By decomposing (2.12) into components and basis vectors, we can view the tangent vector $X_x$ as one giant tuple

$$X_x = \left( \frac{\partial}{\partial u^1}\Big|_x, \ldots, \frac{\partial}{\partial u^d}\Big|_x, X_x^1, \ldots, X_x^d \right) = (\omega(x), X(x)) \in F\mathcal{M} \times \mathbb{R}^d. \tag{2.13}$$

Another choice of coordinate chart produces another local frame $\omega' : U' \to F\mathcal{M}$ and another decomposition (2.12), assuming that $x \in U'$. These decompositions are related by a change of basis

$$(\omega'(x), X'(x)) = (\omega(x)B(x), B(x)^{-1}X(x)), \tag{2.14}$$

for some $B(x) \in GL(d, \mathbb{R})$. Now observe that (2.14) is of the form

$$(p', v') = (p \triangleleft k, \rho(k)^{-1} v), \tag{2.15}$$

where $p, p' \in F\mathcal{M}$, $v, v' \in \mathbb{R}^d$, $k \in GL(d, \mathbb{R})$, the right-action $\triangleleft$ is right-multiplication, and $(\rho, \mathbb{R}^d)$ is the standard representation $\rho(k) = k$ of $GL(d, \mathbb{R})$. The basis-dependent description (2.12) of $X_x$ thus resemble the pairs $(p, v)$ in the construction of associated bundles. Passing to the quotient $(F\mathcal{M} \times_\rho \mathbb{R}^d)/\sim$ instead gives a basis-*in*dependent description of $X_x$, since it identifies all possible decompositions (2.12) in all possible bases. That is, the tangent bundle is isomorphic to $F\mathcal{M} \times_\rho \mathbb{R}^d$. ■

Equivariant neural networks use the language of principal and associated bundles. In the remainder of this subsection, let $E_\rho = P \times_\rho V_\rho$ and $E_\sigma = P \times_\sigma V_\sigma$ be associated bundles, given a principal bundle $\pi : P \to \mathcal{M}$ over a smooth manifold $\mathcal{M}$. Further let $\Gamma_c(E_\rho)$ and $\Gamma_c(E_\sigma)$ be the vector spaces of compactly supported continuous sections of $E_\rho$ and $E_\sigma$, respectively.

**Definition 3.** A *data point* is a section $s \in \Gamma_c(E_\rho)$.

**Remark 3.** Even though data is typically real-valued, we primarily consider complex representations $(\rho, V_\rho)$ so to simplify the mathematical theory. The harmonic analysis in Section 3.5 especially benefits from this choice.

Our decision to restrict attention to compactly supported sections was also made for mathematical reasons: $G$-equivariant layers are defined in Section 3.4 in terms of an induced representation, which lives on the completion of $\Gamma_c(E_\rho)$ with respect to a certain inner product. This is not a serious restriction from an application viewpoint.

**Definition 4.** A *feature map* is a compactly supported continuous map $f : P \to V_\rho$ that satisfies the transformation property

$$f(p \triangleleft k) = \rho(k)^{-1} f(p), \tag{2.16}$$

for all $p \in P$, $k \in K$. The vector space of such feature maps is denoted $C_c(P; \rho)$

Data points and feature maps are, in a sense, dual to each other: Each data point in $\Gamma_c(E_\rho)$ is of the form

$$s_f(x) = [p, f(p)], \tag{2.17}$$

for a feature map $f \in C_c(P; \rho)$, where $p \in P_x$ is any element of the fiber at $x \in \mathcal{M}$. Note that (2.17) does not depend on the choice of $p$: Given another element $p' \in P_x$, there exists a unique $k \in K$ such that $p' = p \triangleleft k$ and

$$[p', f(p')] = [p \triangleleft k, f(p \triangleleft k)] = [p \triangleleft k, \rho(k^{-1}) f(p)] = [p, f(p)]. \tag{2.18}$$

That is, the equivalence class $[p, f(p)]$ only depends on the basepoint $x$.

**Lemma 2** ([26, §10.12])**.** *The linear map $C_c(P; \rho) \to \Gamma_c(E_\rho)$, $f \mapsto s_f$ is a vector space isomorphism.*

We are almost ready to define general and gauge equivariant layers. Before doing so, however, we must say how gauge transformations $\chi : P \to P$ act on data points. Let $\theta_\chi : P \to K$ be the uniquely defined map satisfying $\chi(p) = p \triangleleft \theta_\chi(p)$ for all $p \in P$, and define the following action on the associated bundle $E_\rho$:

$$\chi \cdot [p, v] = [\chi(p), v] = [p, \rho(\theta_\chi(p))v], \qquad [p, v] \in E_\rho. \tag{2.19}$$

The corresponding action on data points is given by

$$(\chi \cdot s_f)(p) = [p, \rho(\theta_\chi(p)) f(p)] = s_{\rho(\theta_\chi) f}(x), \qquad s_f \in \Gamma_c(E_\rho). \tag{2.20}$$

We distinguish between general layers and more specific gauge equivariant layers, as $G$-equivariant layers in GCNNs will only be a special case of the former.

**Definition 5.** Let $E_\rho = P \times_\rho V_\rho$ and $E_\sigma = P \times_\sigma V_\sigma$ be associated bundles.

(i) A *(linear) layer* is a linear map $\Phi : \Gamma_c(E_\rho) \to \Gamma_c(E_\sigma)$.

(ii) A layer $\Phi$ is *gauge equivariant* if, for all gauge transformations $\chi : P \to P$,

$$\Phi \circ \chi = \chi \circ \Phi. \tag{2.21}$$

In equivariant neural networks, data points are sent through a sequence of layers, which are mixed with non-linear activation functions. Again, we focus on individual layers in this paper, and leave the analysis of equivariant activation functions and multi-layer networks for an upcoming paper [23]. The fiber bundle-theoretic concepts discussed in this part describe two kinds of equivariant neural networks:

(i) Gauge equivariant neural networks, which respect local gauge symmetry and whose layers are gauge equivariant.

(ii) GCNNs, which respect global translation symmetry in homogeneous $G$-spaces $\mathcal{M}$, and whose layers are $G$-equivariant (Definition 9).

**Remark 4.** Our definition of layers is almost identical to the linear maps in [9]. The difference is that we focus on compactly supported sections, whereas [9] use sections that are supported on a single coordinate chart. Also, [9] investigates the structure of their linear maps under additional assumptions of so-called locality, covariance, and weight-sharing. Covariance is analogous to gauge equivariance in our setting.

A consequence of Lemma 2 is that (gauge equivariant) layers $\Phi : \Gamma_c(E_\rho) \to \Gamma_c(E_\sigma)$ induce a unique linear map $\phi : C_c(P;\rho) \to C_c(P;\sigma)$ such that $\Phi s_f = s_{\phi f}$. Writing data points as $s_f = [\cdot, f]$ allows us to also express this relation as $\Phi[\cdot, f] = [\cdot, \phi f]$. We think of $\Phi$ and $\phi$ as two sides of the same coin, and use the name (gauge equivariant) layer for both maps.

$$\begin{array}{ccc} \Gamma_c(E_\rho) & \xrightarrow{\quad\Phi\quad} & \Gamma_c(E_\sigma) \\ \uparrow\downarrow & & \uparrow\downarrow \\ C_c(P;\rho) & \xrightarrow{\quad\phi\quad} & C_c(P;\sigma) \end{array}$$

**Example 2.** Let $T : V_\rho \to V_\sigma$ be a linear transformation and consider the layer

$$\phi : C_c(P;\rho) \to C_c(P;\sigma), \qquad (\phi f)(p) = T\big(f(p)\big), \tag{2.22}$$

for $p \in P$, $f \in C_c(P;\rho)$. Since $f$ and $\phi f$ are feature maps and thereby satisfy (2.16), the linear transformation $T$ must satisfy

$$\sigma(k)T\big(f(p)\big) = T\big(f(p \triangleleft k^{-1})\big) = T\big(\rho(k)f(p)\big), \tag{2.23}$$

for all $k \in K, p \in P, f \in C_c(P;\rho)$. This can be seen to imply that $\sigma \circ T = T \circ \rho$, so $T$ intertwines the representations $\rho$ and $\sigma$. Another way to arrive at this conclusion is to analyze when the corresponding layer

$$\Phi : \Gamma_c(E_\rho) \to \Gamma_c(E_\sigma), \qquad \Phi s_f = [\cdot, \phi f], \tag{2.24}$$

is well-defined.

Now consider a gauge transformation $\chi : P \to P$ and its induced map $\theta_\chi : P \to K$. Because $T$ is an intertwiner,

$$(\Phi \circ \chi)s_f = s_{\phi\rho(\theta_\chi)f} = s_{\sigma(\theta_\chi)\phi f} = (\chi \circ \Phi)s_f, \tag{2.25}$$

hence the layer $\Phi$ is automatically gauge equivariant. ■

As this example illustrates, gauge equivariance is tightly connected to intertwining properties of $\phi$. Rearranging (2.25) gives the following result.

**Lemma 3.** *A general layer $\Phi : \Gamma_c(E_\rho) \to \Gamma_c(E_\sigma)$ is gauge equivariant iff*

$$\phi \circ \rho(\theta_\chi)f = \sigma(\theta_\chi) \circ \phi f, \tag{2.26}$$

*for all gauge transformations $\chi : P \to P$ and all feature maps $f \in C_c(P;\rho)$.*

This concludes our discussion of gauge theory and of equivariant neural networks. The framework for the latter is evidently very general, consisting of layers and non-linear activation functions between data points. There are advantages of working at this level of generality: Ordinary (non-equivariant) neural networks have a multitude of different types of layers, many of them linear. Equivariant analogues of such layers are likely to satisfy either Definition 5(ii) or Definition 9, depending on the relevant type of equivariance. Any result that can be proven using this general framework, will thus be true for many different instances of equivariant neural networks. One example is Theorem 14 below, that characterizes the structure of abstract $G$-equivariant layers in any GCNN.

## 3. $G$-equivariant convolutional neural networks

Recall that GCNNs generalize ordinary CNNs to data points $f : \mathcal{M} \to V$ defined on homogeneous $G$-spaces $\mathcal{M}$. Let us give a brief recap on homogeneous spaces and global symmetry, before moving on to discuss homogeneous vector bundles, sections, and induced representations. We will demonstrate that GCNNs and $G$-equivariant layers (originally defined in [12]) are most naturally understood from the perspective of homogeneous vector bundles. We then use reproducing kernel Hilbert spaces and bandwidth to understand which $G$-equivariant layers are expressible as convolutional layers.

### 3.1. **Homogeneous spaces.**

**Definition 6.** Let $G$ be a Lie group. A smooth manifold $\mathcal{M}$ is called a *homogeneous $G$-space* if there exists a smooth, transitive left $G$-action

$$G \times \mathcal{M} \to \mathcal{M}, \qquad (g, x) \mapsto g \cdot x. \tag{3.1}$$

Since the action (3.1) is transitive, we may choose an arbitrary basepoint $x_0 \in \mathcal{M}$ and express any other point $x \in \mathcal{M}$ as $x = g \cdot x_0$ for some $g \in G$. This group element is typically not unique, but observe that

$$g \cdot x_0 = g' \cdot x_0 \quad \iff \quad g^{-1}g' \in H_{x_0}, \tag{3.2}$$

where $H_{x_0} = \{g \in G \mid g \cdot x_0 = x_0\}$ is the isotropy group of $x_0$. In other words, there is a one-to-one correspondence between points $x \in \mathcal{M}$ and left cosets $gH_{x_0} \in G/H_{x_0}$.

**Proposition 4** ([34, Theorem 21.18])**.** *Let $\mathcal{M}$ be a homogeneous $G$-space and choose a basepoint $x_0 \in \mathcal{M}$. The isotropy group $H_{x_0}$ is a closed subgroup of $G$, and the map*

$$F_{x_0} : G/H_{x_0} \to \mathcal{M}, \qquad gH_{x_0} \mapsto g \cdot x_0, \tag{3.3}$$

*is an equivariant diffeomorphism.*

Homogeneous spaces are globally symmetric in the sense that any point $x_0 \in \mathcal{M}$ may be chosen as basepoint. Given another choice of basepoint $x_0' \in \mathcal{M}$, the spaces $G/H_{x_0'} \simeq G/H_{x_0}$ are diffeomorphically related by a translation in $G$ - more precisely, by the composition $F_{x_0}^{-1} \circ F_{x_0'}$. Euclidean space $\mathcal{M} = \mathbb{R}^d$, for example, possesses a global translation symmetry, allowing any point to be considered as origin. Similarly, the rotationally symmetric sphere $\mathcal{M} = S^2$ does not have a unique north pole.

We end this part with the following proposition, which is instrumental in relating homogeneous vector bundles to the equivariance framework in Section 2.2.

**Proposition 5** ([43, §7.5])**.** *Let $G$ be a Lie group and let $H \leq G$ be a closed subgroup. Then the quotient map*

$$q : G \to G/H, \qquad g \mapsto gH, \tag{3.4}$$

*defines a smooth principal $H$-bundle over the homogeneous $G$-space $\mathcal{M} = G/H$.*

### 3.2. **Homogeneous vector bundles.**

Vector bundles may inherit global symmetry from a homogeneous base space; the transitive action $(g, x) \mapsto g \cdot x$ may induce linear maps $E_x \mapsto E_{gx}$ between fibers. Such bundles are naturally called *homogeneous* and, because this symmetry is also encoded in its sections (data points), we will show that homogeneous vector bundles is the natural setting for studying GCNNs.

From this point on, we restrict attention to homogeneous spaces $\mathcal{M} = G/K$ where $G$ is a unimodular Lie group and $K \leq G$ is a compact subgroup. Elements of the homogeneous space is interchangably denoted as $x \in \mathcal{M}$ or $gK \in G/K$.

**Remark 5.** Examples of unimodular Lie groups include all finite, discrete, compact, or abelian Lie groups, the Euclidean groups, and many others. See [19, 20] for details.

**Definition 7** ([47, 5.2.1])**.** Let $\mathcal{M}$ be a homogeneous $G$-space and let $\pi : E \to \mathcal{M}$ be a smooth vector bundle with fibers $E_x$. We say that $E$ is *homogeneous* if there is a smooth left $G$-action $G \times E \to E$ satisfying

$$g \cdot E_x = E_{gx}, \tag{3.5}$$

and such that the induced map $L_{g,x} : E_x \to E_{gx}$ is linear, for all $g \in G, x \in \mathcal{M}$.

**Example 3.** The frame bundle $F\mathcal{M}$ is a homogeneous vector bundle whenever $\mathcal{M}$ is a homogeneous space, and the same is true of any associated bundle $F\mathcal{M} \times_\rho V_\rho$. In particular, the tangent bundle $T\mathcal{M}$ is a homogeneous vector bundle. ■

**Example 4.** If $(\rho, V_\rho)$ is a finite-dimensional $K$-representation, then the associated bundle $E_\rho = G \times_\rho V_\rho$ is a homogeneous vector bundle with respect to the left action

$$g \cdot [g', v] = [gg', v]. \tag{3.6}$$

■

All homogeneous vector bundles $E$ are of the form $G \times_\rho V_\rho$, up to isomorphism. To understand why, consider the fiber $E_K = E_{eK}$ and observe that the restriction of (3.5) to $E_K$ and elements $k \in K$ yields invertible linear maps

$$L_k : E_K \to E_K. \tag{3.7}$$

The defining properties of group actions ensure that $\rho(k) = L_k$ is a finite-dimensional $K$-representation on $E_K$. Moreover, because the linear maps $L_{g,x}$ are isomorphisms, any element $v'$ of any fiber $E_x$ can be obtained as the image $v' = L_{g,K}(v) =: L_g(v)$ for some choices of $g \in q^{-1}(\{x\})$ and $v \in E_K$. The mapping

$$\xi : G \times E_K \to E, \qquad (g, v) \mapsto L_g(v), \tag{3.8}$$

is thus surjective. It is not injective, though, since the relation

$$L_g = L_g \circ L_k \circ L_{k^{-1}} = L_{gk} \circ \rho(k^{-1}), \tag{3.9}$$

implies that $\xi(g, v) = L_g(v) = L_{gk}(\rho(k^{-1})v) = \xi(gk, \rho(k^{-1})v)$ for $k \in K$. However, the same argument shows that $\xi$ is made injective by passing to the quotient $G \times_\rho E_K$.

**Lemma 6** ([47, 5.2.3])**.** *The map*

$$G \times_\rho E_K \to E, \qquad [g, v] \mapsto L_g(v), \tag{3.10}$$

*is an isomorphism of homogeneous vector bundles.*

We now have two perspectives on bundles $G \times_\rho V_\rho$: As bundles associated to the principal bundle $P = G$, and as homogeneous vector bundles (up to isomorphism). The former perspective offers a connection to the framework in Section 2.2, whereas the latter motivates the definition of $G$-equivariant layers in Section 3.4 below.

3.3. **Induced representations.** Let us show the relationship between homogeneous vector bundles and induced representations, which will be an essential ingredient in the definition of $G$-equivariant layers. To this end, let $(\rho, V_\rho)$ be a finite-dimensional unitary $K$-representation and consider the homogeneous vector bundle $E_\rho = G \times_\rho V_\rho$.

We will need inner products on $\Gamma_c(E_\rho)$ and $C_c(G; \rho)$, the former of which is defined using the following unitary structure:

**Lemma 7** ([47, 5.2.7])**.** *The unitary structure*

$$\langle [g, v], [g, w] \rangle_{gK} := \langle v, w \rangle_\rho, \tag{3.11}$$

*defines a complete inner product on each fiber $E_{gK}$, making $E_\rho$ into a Hilbert bundle with $L_{g,x}$ unitary. This unitary structure is unique in that, if we identify $V_\rho$ with $E_K$ in the canonical manner, then the inner product on $V_\rho$ so induced agrees with $\langle\ ,\ \rangle_\rho$.*

We also need the following measure on $G/K$:

**Theorem 8** (Quotient Integral Formula [14, §1.5])**.** *There is a unique $G$-invariant, nonzero Radon measure* $\mathrm{d}x$ *on $G/K$ such that the following* quotient integral formula *holds for every $f \in C_c(G)$:*

$$\int_G f(g)\ \mathrm{d}g = \int_{G/K} \int_K f(xk)\ \mathrm{d}k\ \mathrm{d}x. \tag{3.12}$$

Using these two ingredients, we make $\Gamma_c(E_\rho)$ into a pre-Hilbert space with respect to the inner product

$$\langle s, s' \rangle_{L^2(E_\rho)} := \int_{G/K} \langle s(x), s'(x) \rangle_x\ \mathrm{d}x, \qquad s, s' \in \Gamma_c(E_\rho), \tag{3.13}$$

and we denote its completion $L^2(E_\rho)$. Similarly, $C_c(G;\rho)$ is a pre-Hilbert space with respect to the inner product

$$\langle f, f' \rangle_{L^2(G;\rho)} = \int_G \langle f(g), f'(g) \rangle_\rho\ \mathrm{d}g, \qquad f, f' \in C_c(G;\rho), \tag{3.14}$$

the completion of which is denoted $L^2(G;\rho)$.

**Definition 8.** The $G$-representations

$$\mathrm{ind}_K^G \rho(g) : L^2(E_\rho) \to L^2(E_\rho), \qquad (\mathrm{ind}_K^G \rho(g) s)(x) \ = g \cdot s(g^{-1}x), \tag{3.15}$$

$$\mathrm{Ind}_K^G \rho(g) : L^2(G;\rho) \to L^2(G;\rho), \qquad (\mathrm{Ind}_K^G \rho(g) f)(g') = f(g^{-1}g'). \tag{3.16}$$

are called *induced representations*, or *representations induced by $\rho$.*

Both $\mathrm{ind}_K^G \rho$ and $\mathrm{Ind}_K^G \rho$ are unitary [47, 5.3.2] and may be identified:

**Lemma 9.** *The induced representations* $\mathrm{ind}_K^G(\rho)$, $\mathrm{Ind}_K^G(\rho)$ *are unitarily equivalent.*

*Proof.* This is [47, 5.3.4], but let us write down a proof for clarity. First observe that the isomorphism $C_c(G;\rho) \to \Gamma_c(E_\rho)$, $f \mapsto s_f$ is unitary, which follows by combining the quotient integral formula (3.12), the unitarity of $\rho$, and the compactness of $K$: For all $f, f' \in C_c(G;\rho)$, the map $g \mapsto \langle f(g), f'(g) \rangle_\rho$ lies in $C_c(G)$ and so

$$\begin{aligned} \langle f, f' \rangle_{L^2(G;\rho)} &= \int_{G/K} \int_K \langle f(xk), f'(xk) \rangle_\rho\ \mathrm{d}k\ \mathrm{d}x \\ &= \int_{G/K} \langle f(x), f'(x) \rangle_\rho\ \mathrm{d}x \\ &= \int_{G/K} \langle [x, f(x)], [x, f'(x)] \rangle_x\ \mathrm{d}x = \langle s_f, s_{f'} \rangle_{L^2(E_\rho)}. \end{aligned} \tag{3.17}$$

The same map $f \mapsto s_f$ satisfies

$$\left(\mathrm{ind}_K^G(\rho) s_f\right)(x) = g \cdot s_f(g^{-1}x) = [x, f(g^{-1}x)] = s_{\mathrm{Ind}_K^G(\rho) f}(x), \tag{3.18}$$

so it extends to a unitary isomorphism $L^2(G;\rho) \to L^2(E_\rho)$ intertwining the induced representations. □

To gain a better understanding of the induced representations, consider the Bochner space $L^2(G, V)$, the space of square-integrable functions $f : G \to V$ that take values in a finite-dimensional Hilbert space $V$. It is itself a Hilbert space with inner product

$$\langle f, f' \rangle_{L^2(G,V)} = \int_G \langle f(g), f'(g) \rangle_V\ \mathrm{d}g'. \tag{3.19}$$

The induced representation $(\mathrm{Ind}_K^G \rho, L^2(G;\rho))$ is nothing but the restriction of the left regular representation $\Lambda$ on $L^2(G, V_\rho)$ to a closed, invariant subspace. Furthermore, $\Lambda$ is intimately related to the left regular representation $\lambda$ on $L^2(G)$, as the following lemma shows. The proof of this lemma is a short calculation.

**Lemma 10.** *Let $V$ be a finite-dimensional Hilbert space and equip $L^2(G) \otimes V$ with the tensor product inner product. Then the natural unitary isomorphism*

$$\begin{aligned} A : L^2(G) \otimes V &\to L^2(G, V) \\ f \otimes v &\mapsto fv \end{aligned} \tag{3.20}$$

*intertwines $\lambda \otimes \mathrm{Id}_V$ with $\Lambda$.*

This lemma also shows that, if we choose an orthonormal basis $e_1, \dots, e_{\dim V} \in V$, elements of $L^2(G, V)$ are simply linear combinations $f = \sum_i f^i e_i$ with component functions $f^i \in L^2(G)$. We use this fact in some calculations of vector-valued integrals, and the component functions will also be important in Section 3.5.

3.4. **$G$-equivariant and convolutional layers.** Given a homogeneous $G$-space $\mathcal{M}$, we observed that vector bundles $\pi : E \to \mathcal{M}$ may inherit the global symmetry of $\mathcal{M}$. We took a closer look at such homogeneous vector bundles and found that they are isomorphic to associated bundles $G \times_\rho V_\rho$, and therefore fit within the equivariance framework of Section 2.2 . We also saw how the global symmetry of $\mathcal{M}$ is encoded in data points and feature maps via induced representations, and we want $G$-equivariant layers to preserve this global symmetry.

Consider homogeneous vector bundles $E_\rho = G \times_\rho V_\rho$ and $E_\sigma = G \times_\sigma V_\sigma$, and recall Definition 5 of layers as general linear maps $\Phi : \Gamma_c(E_\rho) \to \Gamma_c(E_\sigma)$. We are mainly interested in *bounded* layers from an application point of view, and we can make this restriction now that the domain and codomain are normed spaces. Furthermore, any bounded layer can be uniquely extended to a bounded linear map

$$\Phi : L^2(E_\rho) \to L^2(E_\sigma), \tag{3.21}$$

and we assume this extension has already been made.

**Definition 9.** A bounded linear map $\Phi : L^2(E_\rho) \to L^2(E_\sigma)$ is called a *$G$-equivariant layer* if it intertwines the induced representations:

$$\Phi \circ \mathrm{ind}_K^G \rho = \mathrm{ind}_K^G \sigma \circ \Phi. \tag{3.22}$$

That is, $G$-equivariant layers are elements $\Phi \in \mathrm{Hom}_G(L^2(E_\rho), L^2(E_\sigma))$.

**Remark 6.** We could also have defined $G$-equivariant layers as bounded linear maps $\phi : L^2(G;\rho) \to L^2(G;\sigma)$ that intertwine the induced representations:

$$\phi \circ \mathrm{Ind}_K^G \rho = \mathrm{Ind}_K^G \sigma \circ \Phi, \tag{3.23}$$

i.e., elements $\phi \in \mathrm{Hom}_G(L^2(G;\rho), L^2(G,\sigma))$. These definitions are clearly equivalent.

Apart from minor technical differences, Definition 9 coincides with the definition of equivariant maps in [12]. We have thus obtained GCNNs almost directly from the definition of homogeneous vector bundles and a desire for layers to respect the global symmetry. This shows that homogeneous vector bundles is the natural setting for GCNNs.

Let us now define convolutional layers.

**Definition 10.** A *convolutional layer* $L^2(G;\rho) \to L^2(G;\sigma)$ is a bounded operator

$$[\kappa \star f](g) = \int_G \kappa(g^{-1}g') f(g') \, \mathrm{d}g', \qquad f \in L^2(G;\rho), \tag{3.24}$$

with an operator-valued kernel $\kappa : G \to \mathrm{Hom}(V_\rho, V_\sigma)$.

Of course, not any function $\kappa : G \to \mathrm{Hom}(V_\rho, V_\sigma)$ can be chosen as the kernel of a convolutional layer. The kernel must ensure both that (3.24) is bounded and that $\phi f \in L^2(G;\sigma)$ for each $f \in L^2(G;\rho)$. We give a sufficient condition for boundedness in Lemma 12 and the other requirement has been studied in detail in [12, 31].

The next result is an almost immediate consequence of the Fubini-Tonelli theorem.

**Proposition 11.** *The adjoint of* (3.24) *is the integral operator*

$$[f * \kappa^*](g) = \int_G \kappa^*(g'^{-1}g) f(g') \, \mathrm{d}g', \qquad f \in L^2(G;\sigma), \tag{3.25}$$

*where $\kappa^*$ is the pointwise adjoint of $\kappa$. That is, $(\kappa \star \cdot)^* = \cdot * \kappa^*$.*

One way to ensure that the operators (3.24)-(3.25) are bounded, is to put a bound on the kernel matrix elements $\kappa_{ij} : G \to \mathbb{C}$ for any given choice of bases in $V_\rho$, $V_\sigma$.

**Lemma 12.** *The operators* (3.24) *and* (3.25) *are bounded if $\kappa_{ij} \in L^1(G)$ for all $i, j$.*

*Proof.* We need only prove that (3.25) is bounded, its adjoint (3.24) will be bounded as well. Choose bases $e_1, \dots, e_{\dim V_\rho} \in V_\rho$ and $\tilde{e}_1, \dots, \tilde{e}_{\dim V_\sigma} \in V_\sigma$ and observe that, because $L^2(G;\sigma) \subset L^2(G, V_\sigma)$, Lemma 10 enables the decomposition of $f \in L^2(G;\sigma)$ into component functions $f^i \in L^2(G)$:

$$f = \sum_{i=1}^{\dim V_\sigma} f^i \tilde{e}_i. \tag{3.26}$$

To be clear, the kernel $\kappa$ is similarly decomposed into matrix elements $\kappa_{ij} = \langle \tilde{e}_j, \kappa e_i \rangle_\sigma$ and we have $\kappa^*_{ji} = \overline{\kappa_{ij}}$. The integral (3.25) now takes the form

$$[f * \kappa^*](g) = \sum_{j=1}^{\dim V_\rho} \left( \int_G \sum_{i=1}^{\dim V_\sigma} \kappa^*_{ji}(g'^{-1}g) f^i(g) \, \mathrm{d}g \right) e_j, \tag{3.27}$$

so by Young's convolution inequality,

$$\begin{aligned} \|f * \kappa^*\|^2_{L^2(G;\rho)} &\leq \sum_{i,j} \int_G \left| \int_G \kappa^*_{ji}(g'^{-1}g) f^i(g') \, \mathrm{d}g' \right|^2 \mathrm{d}g = \sum_{i,j} \|f^i * \kappa^*_{ji}\|_2^2 \\ &\leq \sum_{i,j} \|\kappa_{ij}\|_1^2 \|f^i\|_2^2 \leq M \sum_i \|f^j\|_2^2 = M \|f\|^2_{L^2(G;\sigma)}, \end{aligned} \tag{3.28}$$

where $M = \sum_{i,j} \|\kappa_{ij}\|_1^2 < \infty$ if $\kappa_{ij} \in L^1(G)$ for all $i, j$. □

We are interested in convolutional layers partly because they are concrete examples of $G$-equivariant layers, which we show next.

**Proposition 13.** *Convolutional layers are $G$-equivariant layers.*

*Proof.* Convolutional layers $\kappa \star \cdot : L^2(G;\rho) \to L^2(G;\sigma)$ are bounded linear operators by definition, so the only thing we need to prove is that $\kappa \star \cdot$ intertwines the induced representations. This follows immediately from left-invariance of the Haar measure: For each $f \in L^2(G;\rho)$ and all $g, h \in G$,

$$\begin{aligned} \big[\kappa \star \big(\mathrm{Ind}_K^G \rho(g) f\big)\big](h) &= \int_G \kappa(h^{-1}g') f(g^{-1}g') \, \mathrm{d}g' \qquad (g' \mapsto gg') \\ &= \int_G \kappa\big((g^{-1}h)^{-1}g'\big) f(g') \, \mathrm{d}g' = [\kappa \star f](g^{-1}h), \end{aligned} \tag{3.29}$$

hence $\big[\kappa \star \mathrm{Ind}_K^G \rho(g) f\big] = \mathrm{Ind}_K^G \sigma(g)[\kappa \star f]$. □

**Example 5.** Let us describe where ordinary CNNs fit in the present context. CNNs represent the case $G = \mathbb{Z}^2$ when $K = \{0\}$ is the trivial subgroup. The corresponding homogeneous space is $G/K = \mathbb{Z}^2/\{0\} = \mathbb{Z}^2$ and the quotient map $q : G \to G/K$ is thus the identity map on $\mathbb{Z}^2$. Its inverse, the identity map $\omega : G/K \to G$, is a globally defined gauge that eliminates the need for gauge equivariance, as we may choose to work exclusively in this one gauge. This is just a reflection of the fact that

$$G = \mathbb{Z}^2 = \mathbb{Z}/\{0\} \times \{0\} = G/K \times K, \tag{3.30}$$

is (obviously) trivial as a principal bundle. Its associated bundles $E_\rho = \mathbb{Z}^2 \times_\rho V_\rho$ are also trivial: partly because the finite-dimensional $K$-representation $\sigma$ must be trivial, and partly because each equivalence class $[g, v]$ only contains a single representative. These reasons are, of course, due to the triviality of $K$.

This is not to say that the equivariant framework of Section 2.2 is uninteresting when dealing with CNNs, or with GCNNs for other homogeneous spaces $\mathcal{M} = G/K$ with $K$ trivial. We saw in Sections 3.2-3.3 how the homogeneity give rise to induced representations, which encode the global symmetry in both data points and feature maps. This is a useful perspective to have, and $G$-equivariant layers are interesting even when the bundles are trivial.

Triviality of the associated bundles, $E_\rho \simeq \mathbb{Z}^2 \times \mathbb{C}^m$ where $m = \dim V_\sigma$,[6] implies that data points and feature maps are general square-integrable functions,

$$L^2(E_\rho) \simeq L^2(\mathbb{Z}^2; \rho) \simeq L^2(\mathbb{Z}^2, \mathbb{C}^m), \tag{3.31}$$

and are thereby extensions of compactly supported functions $f : \mathbb{Z}^2 \to \mathbb{C}^m$. This ties well into the discussion in Section 2.1. Convolutional layers (3.24) reduce to bounded linear operators $L^2(\mathbb{Z}^2, \mathbb{C}^m) \to L^2(\mathbb{Z}^2, \mathbb{C}^n)$ and take the form

$$(\kappa \star f)(x) = \sum_{y \in \mathbb{Z}^2} \kappa(y - x) f(y), \tag{3.32}$$

as the Haar measure on $\mathbb{Z}^2$ is the counting measure. The kernel $\kappa : \mathbb{Z}^2 \to \mathrm{Hom}(\mathbb{C}^m, \mathbb{C}^n)$ is finitely supported in practice, so boundeness of (3.32) is ensured by Lemma 12.

Interestingly, all $\mathbb{Z}^2$-equivariant layers are convolutional layers; there are no other types of $\mathbb{Z}^2$-equivariant layers than (3.32). This is a consequence of Theorem 14 and is proven in Corollary 20 below. ■

For more general groups $G$, it is no longer true that all $G$-equivariant layers are convolutional layers; we give an example of this fact in Example 6. Implementations of GCNNs, however, are usually based on convolutional layers, or on analogous layers in the Fourier domain. What consequences does the restriction to convolutional layers have for the expressivity of GCNNs? Can we tell whether a given $G$-equivariant layer is expressible as a convolutional layer? The answer to this last question, it turns out, requires the following notion of reproducing kernel Hilbert spaces.

**Definition 11.** Let $G$ be a group, let $V$ be a finite-dimensional normed vector space, and let $\mathcal{H}$ be a Hilbert space of functions $G \to V$. Then $\mathcal{H}$ is a *reproducing kernel Hilbert space (RKHS)* if the evaluation operator

$$\mathcal{E}_g : \mathcal{H} \to V, \qquad f \mapsto f(g), \tag{3.33}$$

is bounded for all $g \in G$. Moreover, by *left-invariant RKH subspace* $\mathcal{H} \subseteq L^2(G, V)$ we mean a closed subspace that is both a RKHS and an invariant subspace for the left regular representation $\Lambda$ on $L^2(G, V)$.

[6] Recall that we focus on complex vector bundles, hence the use of $\mathbb{C}^m$ instead of $\mathbb{R}^m$.

**Remark 7.** The term RKHS is typically reserved for the scalar case $V = \mathbb{C}$, when the evaluation operator is a linear functional. Our version would instead be dubbed *vector-valued* RKHS. We see little benefit from distinguishing between these cases, however, so we use the term RKHS all-encompassingly.

The name RKHS is due to the existence of a kernel-type function that reproduces all elements of $\mathcal{H}$. To see how, choose an orthonormal basis $e_1, \dots, e_{\dim V} \in V$ and write elements $v \in V$ as linear combinations $v = \sum_i v^i e_i$. The projection $P_i(v) = v^i$ onto the $i$'th component is always continuous, so the composition $\mathcal{E}_{g,i} := P_i \circ \mathcal{E}_g$ is a continuous linear functional

$$\mathcal{E}_g^i : \mathcal{H} \to \mathbb{C}, \qquad f \mapsto f^i(g), \tag{3.34}$$

for all $g \in G$ and $i = 1, \dots, \dim V$. By the Riesz representation theorem, there are elements $\varphi_{g,i} \in \mathcal{H}$ such that $f^i(g) = \mathcal{E}_{g,i}(f) = \langle f, \varphi_{g,i} \rangle$, hence

$$f(g) = \sum_{i=1}^{\dim V} \langle f, \varphi_{g,i} \rangle e_i. \tag{3.35}$$

Now, if $\mathcal{H} \subseteq L^2(G, V)$ is a left-invariant RKH subspace, expanding the functions $\varphi_{g,i}$ in the orthonormal basis, $\varphi_{g,i} = \sum_j \varphi_{g,i}^j e_j$, yields the formula

$$f(g) = \sum_i \left( \int_G \sum_j \overline{\varphi_{g,i}^j(g')} f^j(g') \, \mathrm{d}g' \right) e_i = \int_G \varphi_g^*(g') f(g') \, \mathrm{d}g', \tag{3.36}$$

where $\varphi_g^*$ is the conjugate transpose of the matrix $(\varphi_g)_i^j = \varphi_{g,i}^j$. By left-invariance,

$$f(g) = \big(\Lambda(g^{-1})f\big)(e) = \int_G \varphi_e^*(g^{-1}g') f(g') \, \mathrm{d}g', \tag{3.37}$$

hence $f \in \mathcal{H}$ is *reproduced* by the operator-valued *kernel* $\varphi_e : G \to \mathrm{Hom}(V)$.

**Remark 8.** The reproducing kernel $\varphi_e$ is unique and thus independent of the choice of basis in $V$. This follows from uniqueness in the Riesz representation theorem.

It is now clear why left-invariant RKH subspaces of $L^2(G, V)$ are relevant when discussing convolutional layers, as the latter are given by integral operators similar to (3.37). In order to show that an abstract $G$-equivariant layer $\phi : L^2(G; \rho) \to L^2(G; \sigma)$ can be written as a convolutional layer, it is almost necessary for it to act in a RKHS:

**Example 6.** The identity operator $\phi : L^2(G; \sigma) \to L^2(G; \sigma)$ is clearly a $G$-equivariant layer regardless of $G$, $K$, $\sigma$, but it is only a convolutional layer if $L^2(G; \sigma)$ is a RKHS. This is because when $\phi$ is the identity, (3.24) becomes the reproducing property

$$f(g) = \int_G \kappa(g^{-1}g') f(g') \, \mathrm{d}g', \qquad f \in L^2(G; \sigma). \tag{3.38}$$

It follows that not every $G$-equivariant layer is a convolutional layer, because $L^2(G; \sigma)$ is not always a RKHS. When $\sigma$ is the trivial representation, for instance, $L^2(G; \sigma)$ reduces to $L^2(G)$ which is not a RKHS when $G$ is nondiscrete [20, Theorem 2.42]. ■

At this point, we know that global symmetry manifests itself in feature maps and data points through the induced representation, and we used this knowledge to define $G$-equivariant layers. We also defined convolutional layers and showed that these are special cases of $G$-equivariant layers, but the converse problem is much more subtle: When can a $G$-equivariant layer be expressed as a convolutional layer? The answer, as we have just seen, is directly related to the concept of RKHS and our next result makes this relation precise. It can be considered our main theorem.

**Theorem 14.** *Let $G$ be a unimodular Lie group, let $K \leq G$ be a compact subgroup, and consider homogeneous vector bundles $E_\rho, E_\sigma$ over $\mathcal{M} = G/K$. Suppose that*

$$\phi : L^2(G;\rho) \to L^2(G;\sigma), \tag{3.39}$$

*is a $G$-equivariant layer. If $\phi$ maps into a left-invariant RKH subspace $\mathcal{H} \subseteq L^2(G;\sigma)$, then $\phi$ is a convolutional layer.*

*Proof.* Fix orthonormal bases in $V_\rho$, $V_\sigma$. For $i = 1, \dots, \dim \sigma$, consider the functionals

$$\mathcal{E}^i : L^2(G;\rho) \to \mathbb{C}, \qquad \mathcal{E}^i(f) = (\phi f)^i(e), \tag{3.40}$$

composing $\phi$ with evaluation at the identity element $e \in G$ and projection onto the $i$'th component. As $\phi$ maps into a left-invariant RKH subspace $\mathcal{H} \subseteq L^2(G;\sigma)$, (3.40) is a bounded linear functional: $|\mathcal{E}^i(f)| \leq \|(\phi f)(e)\|_\sigma \leq \|\phi f\|_{L^2(G;\sigma)} \leq \|\phi\| \|f\|_{L^2(G;\rho)}$. By the Riesz representation theorem, there is a unique $\varphi_i \in L^2(G;\rho)$ such that

$$\mathcal{E}^i(f) = \int_G \langle f(g), \varphi_i(g) \rangle_\rho \, \mathrm{d}g = \int_G \sum_{j=1}^{\dim \rho} f^i(g) \overline{\varphi_i^j(g)} \, \mathrm{d}g, \tag{3.41}$$

and proceeding as in (3.36)-(3.37) with $\kappa := \varphi_e^*$ yields the desired relation

$$(\phi f)(g) = \int_G \kappa(g^{-1}g') f(g') \, \mathrm{d}g'. \tag{3.42}$$

□

**Remark 9.** Theorem 14 is a generalization of [12, Theorem 6.1], which was proven under the assumption that $\phi$ is an integral operator $(\phi f)(g) = \int_G \kappa(g, g') f(g') \, \mathrm{d}g'$.

**Remark 10.** While Theorem 14 is similar in spirit to [28, Theorem 1], there are also some clear differences. For example, we work with unimodular Lie groups whereas [28] use compact groups, but [28, Theorem 1] is also stronger in this case as there is no criterion on the layer. Another difference is that [28] analyzes the whole network structure while we focus on individual layers. We also assume that the homogeneous space $G/K$ is the same before and after each layer, in constrast to [28].

In the special case of single-layer networks with compact $G$, [28, Theorem 1] states that any $G$-equivariant layer is a convolutional layer. Example 6 seems to contradict this statement when $G$ is non-discrete compact. This conflict is possibly due to minor technical differences in the assumptions on layers and data points, but we have not identified the precise cause.

We end this section with a result that could simplify the numerical computations of convolutional layers, as integrals over $G/K$ are sometimes easier to compute than integrals over $G$. For example when $G = SO(3)$, $K = SO(2)$, and $G/K \simeq S^2$. This result is similar to the generalized convolutions described in [28, Section 4.1]

**Corollary 15.** *Let $\phi : L^2(G;\rho) \to L^2(G;\sigma)$ be as in Theorem 14 and let $\kappa$ be the kernel of the resulting convolutional layer* (3.42)*. Then*

$$(\phi f)(g) = \int_{G/K} \kappa(g^{-1}x) f(x) \, \mathrm{d}x. \tag{3.43}$$

*Proof.* In the proof of Theorem 14, we constructed the kernel $\kappa$ from the components of $\varphi_i \in L^2(G;\rho)$, and unitarity of $\rho$ clearly implies that the expression $\langle f(x), \varphi_i(x) \rangle_\rho$

is well-defined. We may therefore use the unitary structure (3.11) to get the following relation for all component functions $(\phi f)^i$ and all $g \in G$:

$$(\phi f)^i(g) = \langle f, \operatorname{Ind}_K^G \rho(g)\varphi_i \rangle_{L^2(G;\rho)} = \langle s_f, s_{\operatorname{Ind}_K^G \rho(g)\varphi_i} \rangle_{L^2(E_\rho)} \tag{3.44}$$

$$= \int_{G/K} \langle s_f(x), s_{\operatorname{Ind}_K^G \rho(g)\varphi_i}(x) \rangle_x \, \mathrm{d}x \tag{3.45}$$

$$= \int_{G/K} \langle f(x), \operatorname{Ind}_K^G \rho(g)\varphi_i(x) \rangle_\rho \, \mathrm{d}x \tag{3.46}$$

$$= \int_{G/K} \sum_{j=1}^{\dim \rho} \overline{\varphi_i^j(g^{-1}x)} f^j(x) \, \mathrm{d}x. \tag{3.47}$$

We now obtain (3.43) by reconstructing $\kappa$ from its components $\kappa_{ij} = \overline{\varphi_i^j}$. □

3.5. **RKHS and bandlimited functions.** The strength of Theorem 14 naturally depends on how common left-invariant RKH subspaces of $L^2(G;\sigma)$ are. Our analysis of $G$-equivariant layers would not be complete without a discussion on this topic.

Let us proceed by investigating when the component functions $f^i$ of $f \in L^2(G;\sigma)$ are contained in a left-invariant RKH subspace $\mathcal{H} \subset L^2(G)$; these subspaces have been fully characterized when the unimodular Lie group $G$ is of *type I* [8, 20]. The unitary equivalence (3.20) then ensures that $A(\mathcal{H} \otimes V_\sigma) \subset L^2(G, V_\sigma)$ is a left-invariant RKH subspace, and so is the closed subspace

$$A(\mathcal{H} \otimes V_\sigma) \cap L^2(G;\sigma) \subset L^2(G;\sigma). \tag{3.48}$$

**Remark 11.** Groups of type I are, in a sense, groups with manageable representation theory. They include the most common groups, such as all finite, discrete, compact, or abelian groups, the Euclidean groups and many other groups. In particular, there is a considerable overlap between type I groups and the unimodular Lie groups that we already consider. See [19, 20] for more details.

**Remark 12.** While $\rho, \sigma$ still denote finite-dimensional unitary representations of $K$, we reserve the letter $\gamma$ for elements of the unitary dual $\widehat{G}$, i.e., the space of equivalence classes of unitary representations. Specific representatives of $\gamma$ are written as $(\pi_\gamma, V_\gamma)$, and note that $V_\gamma$ need not be finite-dimensional unless $G$ is compact. The unimodular Lie group $G$ is assumed to be of type I throughout this section.

**Proposition 16** ([20, Proposition 2.40])**.** *Let $\mathcal{H} \subseteq L^2(G)$ be a left-invariant RKH subspace. The kernel $\varphi \in \mathcal{H}$ is then a self-adjoint convolution idempotent,*[7] *and*

$$\mathcal{H} = L^2(G) * \varphi = \left\{ f * \varphi \;\middle|\; f \in L^2(G) \right\} \subset C(G). \tag{3.49}$$

*Conversely, if $\varphi \in L^2(G)$ is a self-adjoint convolution idempotent, then $\mathcal{H} = L^2(G) * \varphi$ is a left-invariant RKH subspace of $L^2(G)$.*

**Example 7.** Consider the real line $G = \mathbb{R}$ and suppose $\mathcal{H} \subseteq L^2(\mathbb{R})$ is a left-invariant RKH subspace with kernel $\varphi \in \mathcal{H}$. The calculation in (3.37) with $V = \mathbb{C}$ shows that, for all $f \in \mathcal{H}$,

$$f(x) = \int_{-\infty}^{\infty} \overline{\varphi(y - x)} f(y) \, \mathrm{d}y = \langle f, \lambda(x)\varphi \rangle = (f * \varphi^*)(x). \tag{3.50}$$

Since the regular representation $\lambda$ is continuous, $f$ must be continuous, so $\mathcal{H} \subset C(\mathbb{R})$. Setting $f = \varphi$ shows that the kernel is a self-adjoint convolution idempotent:

$$\varphi = \varphi * \varphi^* = (\varphi * \varphi^*)^* = \varphi^*. \tag{3.51}$$

[7] That is, $\varphi = \varphi * \varphi^* = \varphi^*$ where $\varphi^*(g) := \overline{\varphi(g^{-1})}$ denotes involution.

Combining the Plancherel transform on $L^2(\mathbb{R})$ (see Theorem 17 and Section 3.5.1) with the convolution theorem in Fourier analysis, we observe that, for all $f \in \mathcal{H}$,

$$\hat{f} = \widehat{f * \varphi} = \hat{f}\hat{\varphi}. \tag{3.52}$$

In particular, $\hat{\varphi} = \hat{\varphi}^2$, so $\hat{\varphi}$ is the characteristic function $1_E$ on a subset $E \subset \widehat{\mathbb{R}} \simeq \mathbb{R}$. Inserting $\hat{\varphi} = 1_E$ in (3.52) immediately tells us that $\mathrm{supp}(\hat{f}) \subset E$, so $\mathcal{H}$ is a space of bandlimited functions. Moreover, the set $E$ has finite Lebesgue measure according to the Plancherel theorem: $\mathrm{vol}(E) = \|1_E\|_2^2 = \|\varphi\|_2^2 < \infty$. ■

This example illustrates that any measurable subset $E \subset \mathbb{R}$ with finite Lebesgue measure induces a left-invariant closed RKH subspace

$$\mathcal{H}_E = \left\{ f \in L^2(\mathbb{R}) \;\middle|\; \mathrm{supp}(\hat{f}) \subset E \right\} = L^2(\mathbb{R}) * \varphi_E, \tag{3.53}$$

$\varphi_E$ being the inverse Plancherel transform of $1_E$ [20, 2.63-2.65]. This relation between left-invariant RKH subspaces $\mathcal{H} \subseteq L^2(G)$ and bandlimited functions generalizes to unimodular Lie groups $G$ of type I, although the necessary harmonic analysis becomes significantly more advanced. Going into detail on this rather technical subject would distract from the topic at hand, so we refer curious readers to the relevant literature instead [20]. Let us take the short route of stating a theorem on the direct integral decomposition of the left regular representation $\lambda$ and its commutant

$$\lambda(G)' = \left\{ T \in \mathcal{B}(L^2(G)) \;\middle|\; T \circ \lambda(g) = \lambda(g) \circ T \text{ for all } g \in G \right\}, \tag{3.54}$$

and discuss a few consequences of this decomposition, before restricting attention to two important cases where we can be more explicit: Abelian and compact groups.

**Definition 12** ([20, §3.5])**.** The *operator-valued Fourier transform* on $G$ maps each $f \in L^1(G)$ to the family $\mathcal{F}(f) = (\hat{f}(\gamma))_{\gamma \in \widehat{G}}$, where each $\hat{f}(\gamma) \in \mathcal{B}(V_\gamma)$ is a bounded operator given by the Bochner integral

$$\hat{f}(\gamma) = \int_G f(g)\pi_\gamma(g)\, \mathrm{d}g. \tag{3.55}$$

**Theorem 17** ([20, Theorem 3.48])**.** *There is a canonical* Plancherel measure $\nu$ *for the unitary dual* $\widehat{G}$ *with the following properties:*

(a) $\mathcal{F}$ *extends to a unitary operator*

$$\mathcal{P} : L^2(G) \to \int_{\widehat{G}}^{\oplus} V_\gamma \otimes V_\gamma \, \mathrm{d}\nu(\gamma), \tag{3.56}$$

*called the* Plancherel transform *of* $G$*.*

(b) $\mathcal{P}$ *implements the following unitary equivalences:*

$$\lambda \simeq \int_{\widehat{G}}^{\oplus} \pi_\gamma \otimes \mathrm{Id} \;\, \mathrm{d}\nu(\gamma) \tag{3.57}$$

$$\lambda(G)' \simeq \int_{\widehat{G}}^{\oplus} \mathrm{Id} \otimes \mathcal{B}(V_\gamma) \, \mathrm{d}\nu(\gamma) \tag{3.58}$$

Observe that if $\mathcal{H} \subseteq L^2(G)$ is a left-invariant closed subspace, then the projection $P : L^2(G) \to \mathcal{H}$ commutes with the left-regular representation and is thus an element of the commutant $\lambda(G)'$. It therefore has a direct integral decomposition

$$P = \int_{\widehat{G}}^{\oplus} \mathrm{Id} \otimes \hat{P}_\gamma \, \mathrm{d}\nu(\gamma), \tag{3.59}$$

where $\hat{P}_\gamma \in \mathcal{B}(V_\gamma)$ for each $\gamma \in \widehat{G}$.

**Theorem 18** ([20, Theorem 4.22, Proposition 2.40])**.** *Suppose that $\mathcal{H} \subseteq L^2(G)$ is a left-invariant closed subspace with equation* (3.59) *denoting the projection onto $\mathcal{H}$. Then $\mathcal{H}$ is a RKHS iff*

$$\int_{\widehat{G}} \operatorname{rank}(\hat{P}_\gamma)\ \mathrm{d}\nu(\gamma) < \infty. \tag{3.60}$$

We interpret this theorem as a bandwidth restriction, similar to Example 7. The integrand in (3.60) is an integer-valued function on $\widehat{G}$, so the integral is finite only if the projection (3.59) is supported on a set $E \subseteq \widehat{G}$ of finite Plancherel measure,

$$\nu\Big(\underbrace{\overline{\big\{\gamma \in \widehat{G} : \hat{P}_\gamma \neq 0\big\}}}_{E}\Big) = \int_{\widehat{G}} 1_E(\gamma)\ \mathrm{d}\nu(\gamma) \leq \int_{\widehat{G}} \operatorname{rank}(\hat{P}_\gamma)\ \mathrm{d}\nu(\gamma). \tag{3.61}$$

That is, the left-invariant RKH subspaces $\mathcal{H} \subseteq L^2(G)$ are precisely those subspaces whose elements are bandlimited on a set $E \subseteq \widehat{G}$, in the sense that, for all $f \in \mathcal{H}$ and each equivalence class $\gamma \notin E$,

$$\hat{f}(\gamma) = \widehat{Pf}(\gamma) = \hat{f}(\gamma) \circ \hat{P}_\gamma = 0. \tag{3.62}$$

**Remark 13.** The second equality in (3.62) is [20, Corollary 4.17].

Bandlimited functions are thus central to the theory of RKHS and, by extension, to the mathematical theory of GCNNs. Indeed, by extending the concept of bandwidth to feature maps, through (3.48), we obtain the following rephrasing of Theorem 14.

**Corollary 19.** *Let $G$ be a unimodular Lie group of type I, let $K \leq G$ be a compact subgroup, and consider homogeneous vector bundles $E_\rho, E_\sigma$ over $G/K$. Suppose that*

$$\phi : L^2(G; \rho) \to L^2(G; \sigma), \tag{3.63}$$

*is a $G$-equivariant layer. If $\phi$ maps into a space of bandlimited functions, then $\phi$ is a convolutional layer.*

**Remark 14.** The relevance of bandwidth for convolutional layers has already been recognized in the case of azimuthally equivariant linear operators on $L^2(S^2)$ [45]. In our setting, these operators translate to certain $G$-equivariant layers

$$\Phi : L^2(E_\rho) \to L^2(E_\sigma), \tag{3.64}$$

when $G = SO(3)$, $K = SO(2)$, and $\rho, \sigma$ are the trivial representation.

**Remark 15.** Some implementations of GCNNs use Fourier transforms and a variant of the convolution theorem $\widehat{f_1 * f_2} = \hat{f}_1 \hat{f}_2$ to compute convolutional layers [16, 29, 45, 48]. Feature maps $f$ are then represented by their Fourier transform $\hat{f}$ which, for numerical reasons, is only approximated up to a finite bandlimit. That is, bandwidth is already being used in implementations.

3.5.1. *Abelian groups.* The irreducible representations $\gamma \in \widehat{G}$ of any abelian group $G$ are 1-dimensional, and may thus be identified with their character $\chi_\gamma = \operatorname{tr} \pi_\gamma$. There are several useful consequences of this fact.

First, the unitary dual $\widehat{G}$ is now the set of continuous homomorphisms $\chi : G \to \mathbb{T}$, where $\mathbb{T}$ is the circle group. This is a locally compact group with respect to pointwise multiplication and, as $\gamma \in \widehat{G}$ is unitary, we may write $\chi_\gamma = e^{i\xi_\gamma}$ where $\xi_\gamma : G \to \mathbb{R}$. The Fourier transform then takes the more familiar form

$$\hat{f}(\gamma) = \int_G f(g) e^{-i\xi_\gamma(g)}\ \mathrm{d}g, \tag{3.65}$$

for $f \in L^1(G) \cap L^2(G)$. Moreover, the Haar measure on $\widehat{G}$ can be made to coincide with the Plancherel measure such that (3.56) becomes a unitary equivalence

$$\mathcal{P} : L^2(G) \to L^2(\widehat{G}). \tag{3.66}$$

Another consequence of the fact that irreducible representations are 1-dimensional, is that the integrand in (3.60) takes values in $\{0, 1\}$ and (3.61) becomes an equality. By the same arguments as in Example 7, we see that the left-invariant RKH subspaces $\mathcal{H} \subset L^2(G)$ are the spaces of bandlimited functions, $\text{supp}(\hat{f}) \subset E$, for subsets $E \subset \widehat{G}$ of finite Haar/Plancherel measure. Also, the kernel $\varphi_E \in \mathcal{H}$ is the inverse Plancherel transform of the characteristic function $1_E$.

**Corollary 20.** *If $G$ is a discrete and abelian group, then any $G$-equivariant layer is a convolutional layer.*

*Proof.* Discrete groups are unimodular Lie group of type I, so we may use results from the current section. We note that the integral (3.60) converges for any left-invariant, closed subspace $\mathcal{H} \subseteq L^2(G)$, as the unitary dual $\widehat{G}$ is compact when $G$ is discrete and abelian [14, Proposition 3.1.5], and because the integrand is bounded. Consequently, $\mathcal{H} = L^2(G)$ is itself a RKHS,[8] and the same is true for both $L^2(G, V_\sigma) \simeq L^2(G) \otimes V_\sigma$ and its closed, left-invariant subspace $L^2(G; \sigma)$, independently of $K \leq G$ and $(\sigma, V_\sigma)$. The result now follows from Theorem 14. □

By setting $G = \mathbb{Z}^2$, Corollary 20 establishes that convolutional layers are the only possible translation equivariant layers in the ordinary CNN setting.

3.5.2. *Compact groups.* When the group $G$ is compact, all irreducible representations are finite-dimensional. Furthermore, the unitary dual $\widehat{G}$ is discrete and the Plancherel measure on $\widehat{G}$ is simply the counting measure. For these reasons, the integral (3.60) reduces to a discrete sum with finite summands, and converges iff $\hat{P}_\gamma = 0$ for all but finitely many $\gamma \in \widehat{G}$.

**Corollary 21.** *If $G$ is finite, then any $G$-equivariant layer is a convolutional layer.*

*Proof.* When $G$ is a finite group, $\widehat{G}$ is also finite [19, Proposition 5.27] and the integral (3.60) reduces to a finite sum. That is, $L^2(G)$ is a RKHS and the result now follows in the same way as Corollary 20. □

## 4. Discussion

In this paper, we have investigated the mathematical foundations of $G$-equivariant convolutional neural networks (GCNNs), which are designed for deep learning tasks exhibiting global symmetry. We presented a basic framework for equivariant neural networks that include both gauge equivariant neural networks and GCNNs as special cases. We also demonstrated how GCNNs can be obtained from homogeneous vector bundles, when $G$ is a unimodular Lie group and $K \leq G$ is a compact subgroup.

In Theorem 14, we gave a precise criterion for when a given $G$-equivariant layer is, in fact, a convolutional layer. This criterion uses reproducing kernel Hilbert spaces (RKHS) and cannot be circumvented, as shown in Example 6. After discussing the relation between RKHS and bandwidth, we were able to reformulate Theorem 14 to get an analogous bandlimit-criterion in Corollary 19. In Corollaries 20-21, we showed that the criterion is automatically satisfied when $G$ is discrete abelian or finite, hence all $G$-equivariant layers are convolutional layers for these groups.

[8] In fact, the kernel is simply the Kronecker delta $\varphi(g) = \delta_{g,e}$.

One limitation of the current paper, compared to [12, 28], is that the homogeneous space $G/K$ does not change between layers. This restriction was made in order to limit the scope of our analysis, and the same goes for our restriction to unimodular Lie groups. It would be interesting to go beyond these restrictions in the future.

I am grateful for the support from my research group: Oscar Carlsson, Jan Gerken, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and last but not least, my advisor Daniel Persson. Thanks also to David Müller for the interesting discussions on gauge equivariance in physics. This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

## References

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *International Conference on Engineering and Technology (ICET)*. IEEE, 2017.

[2] E. Alpaydin. *Introduction to machine learning*. MIT Press, 2014.

[3] D. L. Boyda, M. N. Chernodub, N. V. Gerasimeniuk, V. A. Goy, S. D. Liubimov, and A. V. Molochkov. Finding the deconfinement temperature in lattice yang-mills theories from outside the scaling window with machine learning. *Physical Review D*, 2021.

[4] A. Brock, S. De, S. L. Smith, and K. Simonyan. High-performance large-scale image recognition without normalization. *arXiv preprint arXiv:2102.06171*, 2021.

[5] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 2017.

[6] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[7] W. Cao, Z. Yan, Z. He, and Z. He. A comprehensive survey on geometric deep learning. *IEEE Access*, 2020.

[8] A. L. Carey. Group representations in reproducing kernel hilbert spaces. *Reports on Mathematical Physics*, 1978.

[9] M. Cheng, V. Anagiannis, M. Weiler, P. de Haan, T. Cohen, and M. Welling. Covariance in physics and convolutional neural networks. *arXiv preprint arXiv:1906.02481*, 2019.

[10] T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*. PMLR, 2016.

[11] T. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

[12] T. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant cnns on homogeneous spaces. *arXiv preprint arXiv:1906.02481*, 2018.

[13] T. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International Conference on Machine Learning*. PMLR, 2019.

[14] A. Deitmar and S. Echterhoff. *Principles of harmonic analysis*. Springer International Publishing, 2014.

[15] C. Esteves. Theoretical aspects of group equivariant neural networks. *arXiv preprint arXiv:2004.05154*, 2020.

[16] C. Esteves, C.e Allen-Blanchette, A. Makadia, and K. Daniilidis. 3D object classification and retrieval with spherical cnns. *arXiv preprint arXiv:1711.06721*, 2017.

[17] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] M. Favoni, A. Ipp, D. Müller, and D. Schuh. Lattice gauge equivariant convolutional neural networks. *arXiv preprint arXiv:2012.12901*, 2020.

[19] G. B. Folland. *A course in abstract harmonic analysis*. CRC press, 2016.

[20] H. Führ. *Abstract harmonic analysis of continuous wavelet transforms*. Springer Science & Business Media, 2005.

[21] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*. Springer, 1982.

[22] C. Gattringer and C. Lang. *Quantum chromodynamics on the lattice: an introductory presentation*. Springer Science & Business Media, 2009.

[23] J. E. Gerken, J. Aronsson, O. Carlsson, H. Linander, F. Ohlsson, C. Petersson, and D. Persson. Geometric deep learning and equivariant neural networks. Forthcoming.

[24] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016. URL `http://www.deeplearningbook.org`.

[25] C. Jia, Y. Yang, Y. Xia, Y-T. Chen, Z. Parekh, H. Pham, Q.V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[26] I. Kolár, P. W. Michor, and J. Slovák. *Natural operations in differential geometry*. Springer Science & Business Media, 2013.

[27] O. Koller, S. Zargaran, H. Ney, and R. Bowden. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *International Journal of Computer Vision*, 2018.

[28] R. Kondor and S. Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*. PMLR, 2018.

[29] R. Kondor, Z. Lin, and S. Trivedi. Clebsch-gordan nets: a fully fourier space spherical convolutional neural network. *arXiv preprint arXiv:1806.09231*, 2018.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

[31] L. Lang and M. Weiler. A wigner-eckart theorem for group equivariant convolution kernels. *arXiv preprint arXiv:2010.10952*, 2020.

[32] S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 1997.

[33] Y. LeCun, L.D. Jackel, L. Bottou, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U.A. Muller, E. Sackinger, P. Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 1995.

[34] J. M. Lee. *Introduction to smooth manifolds*. Springer, 2013.

[35] J. Li, Y. Bi, and G.H. Lee. Discrete rotation equivariance for point cloud recognition. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[36] D. Luo, G. Carleo, B. Clark, and J. Stokes. Gauge equivariant neural networks for quantum lattice gauge theories. *Bulletin of the American Physical Society*,

2021.

[37] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 2017.

[38] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[39] D. Müller. Yang-mills theory, lattice gauge theory and simulations, 2019. URL `https://www.jku.at/en/institut-of-analysis/research/guest-lecture-series/`.

[40] M. Nakahara. *Geometry, topology and physics*. CRC press, 2003.

[41] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters - improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[42] S. Song, H. Huang, and T. Ruan. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications*, 2019.

[43] N. Steenrod. *The topology of fibre bundles*. Princeton University Press, 1960.

[44] M. Tan and Q.V. Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.

[45] C. Toft, G. Bökman, and F. Kahl. Azimuthal rotational equivariance in spherical cnns. 2021. URL `https://openreview.net/forum?id=sp3Z1jiS2vn`.

[46] B.S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018.

[47] N. R. Wallach. *Harmonic analysis on homogeneous spaces*. M. Dekker, 2018.

[48] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *arXiv preprint arXiv:1807.02547*, 2018.

[49] D. Worrall and G. Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[50] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[51] J. Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 2017.

[52] Y. Xu, Z. Jia, L-B. Wang, Y. Ai, F. Zhang, M. Lai, I Eric, and C. Chang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics*, 2017.

[53] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Chalmers University of Technology, Department of Mathematical Sciences, SE-412 96 Gothenburg, Sweden

*Email address*: `jimmyar@chalmers.se`